6. Appendix

6.1. Closed-form computation of uncertainty measures & Uncertainty attacks

Dirichlet-based uncertainty models allow to compute several uncertainty measures in closed form (see (Malinin & Gales, 2018a) for a derivation). As proposed by (Malinin & Gales, 2018a), we use precision m_{α_0}, differential entropy m_{diffE} and mutual information m_{MI} to estimate uncertainty on predictions.

The differential entropy m_{diffE} of a DBU model reaches its maximum value for equally probable categorical distributions and thus, a on flat Dirichlet distribution. It is a measure for distributional uncertainty and expected to be low on ID data, but high on OOD data.

$$\begin{aligned} m_{\text{diffE}} = &\sum_{c}^{K} \ln \Gamma(\alpha_c) - \ln \Gamma(\alpha_0) \\ &- \sum_{c}^{K} (\alpha_c - 1) \cdot (\Psi(\alpha_c) - \Psi(\alpha_0)) \end{aligned} \tag{7}$$

where α are the parameters of the Dirichlet-distribution, Γ is the Gamma function and Ψ is the Digamma function.

The mutual information m_{MI} is the difference between the total uncertainty (entropy of the expected distribution) and the expected uncertainty on the data (expected entropy of the distribution). This uncertainty is expected to be low on ID data and high on OOD data.

$$m_{\text{MI}} = -\sum_{c=1}^{K} \frac{\alpha_c}{\alpha_0} \left(\ln \frac{\alpha_c}{\alpha_0} - \Psi(\alpha_c + 1) + \Psi(\alpha_0 + 1) \right) \tag{8}$$

Furthermore, we use the precision α_0 to measure uncertainty, which is expected to be high on ID data and low on OOD data.

$$m_{\alpha_0} = \alpha_0 = \sum_{c=1}^{K} \alpha_c \tag{9}$$

As these uncertainty measures are computed in closed form and it is possible to obtain their gradients, we use them (i.e. m_{diffE}, m_{MI}, m_{α_0}) are target function of our uncertainty attacks. Changing the attacked target function allows us to use a wide range of gradient-based attacks such as FGSM attacks, PGD attacks, but also more sophisticated attacks such as Carlini-Wagner attacks.

6.2. Details of the Experimental setup

Models. We trained all models with a similar based architecture. We used namely 3 linear layers for vector data sets, 3 convolutional layers with size of 5 + 3 linear layers for MNIST and the VGG16 (Simonyan & Zisserman, 2015) architecture with batch normalization for CIFAR10. All the implementation are performed using Pytorch (Paszke et al., 2019). We optimized all models using Adam optimizer. We performed early stopping by checking for loss improvement every 2 epochs and a patience of 10. The models were trained on GPUs (1 TB SSD).

We performed a grid-search for hyper-parameters for all models. The learning rate grid search was done in $[1e^{-5}, 1e^{-3}]$. For PostNet, we used Radial Flows with a depth of 6 and a latent space equal to 6. Further, we performed a grid search for the regularizing factor in $[1e^{-7}, 1e^{-4}]$. For PriorNet, we performed a grid search for the OOD loss weight in $[1, 10]$. For DDNet, we distilled the knowledge of 5 neural networks after a grid search in $[2, 5, 10, 20]$ neural networks. Note that it already implied a significant overhead at training compare to other models.

Metrics. For all experiments, we focused on using AUC-PR scores since it is well suited to imbalance tasks (Saito & Rehmsmeier, 2015) while bringing theoretically similar information than AUC-ROC scores (Davis & Goadrich, 2006). We scaled all scores from $[0, 1]$ to $[0, 100]$. All results are average over 5 training runs using the best hyper-parameters found after the grid search.

Data sets. For vector data sets, we use 5 different random splits to train all models. We split the data in training, validation and test sets (60%, 20%, 20%).

We use the segment vector data set (Dua & Graff, 2017), where the goal is to classify areas of images into 7 classes (window, foliage, grass, brickface, path, cement, sky). We remove class window from ID training data to provide OOD training data to PriorNet. Further, We remove the class 'sky' from training and instead use it as the OOD data set for OOD detection experiments. Each input is composed of 18 attributes describing the image area. The data set contains $2,310$ samples in total.

We further use the Sensorless Drive vector data set (Dua & Graff, 2017), where the goal is to classify extracted motor current measurements into 11 different classes. We remove class 9 from ID training data to provide OOD training data to PriorNet. We remove classes 10 and 11 from training and use them as the OOD dataset for OOD detection experiments. Each input is composed of 49 attributes describing motor behaviour. The data set contains $58,509$ samples in total.

Additionally, we use the MNIST image data set (LeCun & Cortes, 2010) where the goal is to classify pictures of hand-drawn digits into 10 classes (from digit 0 to digit 9). Each input is composed of a $1 \times 28 \times 28$ tensor. The data set contains $70,000$ samples. For OOD detection experiments, we use FashionMNIST (Xiao et al., 2017) and KMNIST

(Clanuwat et al., 2018) containing images of Japanese characters and images of clothes, respectively. FashionMNIST was used as training OOD for PriorNet while KMNIST is used as OOD at test time.

Finally, we use the CIFAR10 image data set (Krizhevsky et al., 2009) where the goal is to classify a picture of objects into 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). Each input is a $3 \times 32 \times 32$ tensor. The data set contains $60,000$ samples. For OOD detection experiments, we use street view house numbers (SVHN) (Netzer et al., 2011) and CIFAR100 (Krizhevsky et al., 2009) containing images of numbers and objects respectively. CIFAR100 was used as training OOD for PriorNet while SVHN is used as OOD at test time.

Perturbations. For all label and uncertainty attacks, we used Fast Gradient Sign Methods and Project Gradient Descent. We tried 6 different attack radii $[0.0, 0.1, 0.2, 0.5, 1.0, 2.0, 4.0]$. These radii operate on the input space after data normalization. We bound perturbations by L_∞-norm or by L_2-norm, with

$$L_\infty(x) = \max_{i=1,\dots,D} |x_i| \quad \text{and} \quad L_2(x) = (\sum_{i=1}^{D} x_i^2)^{0.5}. \tag{10}$$

For L_∞-norm it is obvious how to relate perturbation size ε with perturbed input images, because all inputs are standardized such that the values of their features are between 0 and 1. A perturbation of size $\varepsilon = 0$ corresponds to the original input, while a perturbation of size $\varepsilon = 1$ corresponds to the whole input space and allows to change all features to any value.

For L_2-norm the relation between perturbation size ε and perturbed input images is less obvious. To justify our choice for ε w.r.t. this norm, we relate perturbations size ε_2 corresponding to L_2-norm with perturbations size ε_∞ corresponding to L_∞-norm. First, we compute ε_2, such that the L_2-norm is the smallest super-set of the L_∞-norm. Let us consider a perturbation of ε_∞. The largest L_2-norm would be obtained if each feature is perturbed by ε_∞. Thus, perturbation ε_2, such that L_2 encloses L_∞ is $\varepsilon_2 = (\sum_{i=1}^{D} \varepsilon_\infty^2)^{0.5} = \sqrt{D}\varepsilon_\infty$. For the MNIST-data set, with $D = 28 \times 28$ input features L_2-norm with $\varepsilon_2 = 28$ encloses L_∞-norm with $\varepsilon_\infty = 1$.

Alternatively, ε_2 can be computes such that the volume spanned by L_2-norm is equivalent to the one spanned by L_∞-norm. Using that the volume spanned by L_∞-norm is ε_∞^D and the volume spanned by L_2-norm is $\frac{\pi^{0.5D}\varepsilon_2^D}{\Gamma(0.5D+1)}$ (where Γ is the Gamma-function), we obtain volume equivalence if $\varepsilon_2 = \Gamma(0.5D+1)^{\frac{1}{D}}\sqrt{\pi}\varepsilon_\infty$. For the MNIST-data set, with $D = 28 \times 28$ input features L_2-norm with $\varepsilon_2 \approx 21.39$ is volume equivalent to L_∞-norm with $\varepsilon_\infty = 1$.

6.3. Additional Experiments

Table 8 and 9 illustrate that no DBU model maintains high accuracy under gradient-based label attacks. Accuracy under PGD attacks decreases more than under FGSM attacks, since PGD is stronger. Interestingly Noise attacks achieve also good performances with increasing Noise standard deviation. Note that the attack is not constraint to be with a given radius for Noise attacks.

Table 8. Accuracy under PGD label attacks.

Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST							CIFAR10						
PostNet	**99.4**	**99.2**	**98.8**	96.8	89.6	53.8	13.0	89.5	73.5	51.7	13.2	2.2	0.8	0.3
PriorNet	99.3	99.1	**98.8**	97.4	**93.9**	**75.3**	4.8	88.2	**77.8**	**68.4**	**54.0**	**37.9**	**17.5**	**5.1**
DDNet	**99.4**	99.1	**98.8**	**97.5**	91.6	48.8	0.2	86.1	73.9	59.1	20.5	1.5	0.0	0.0
EvNet	99.2	98.9	98.4	96.8	92.4	73.1	**40.9**	**89.8**	71.7	48.8	11.5	2.7	1.5	0.4
	Sensorless							Segment						
PostNet	98.3	13.1	6.4	4.0	**7.0**	**9.8**	**11.3**	98.9	82.8	**50.1**	**19.2**	**8.8**	**5.1**	**8.6**
PriorNet	**99.3**	16.5	5.6	1.2	0.4	0.2	1.6	**99.5**	90.7	47.6	7.8	0.2	0.0	0.4
DDNet	**99.3**	12.4	2.4	0.6	0.3	0.1	0.1	99.2	**90.8**	45.7	6.9	0.0	0.0	0.0
EvNet	99.0	**35.3**	**22.3**	**11.2**	**7.0**	5.2	4.0	99.3	91.8	54.0	10.3	0.8	0.5	0.6

Table 9. Accuracy under FGSM label attacks.

Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST							CIFAR10						
PostNet	**99.4**	**99.2**	**98.9**	97.7	95.2	**90.1**	**79.2**	89.5	72.3	54.9	31.2	21.0	16.8	15.6
PriorNet	99.3	99.1	**98.9**	97.7	**95.8**	93.2	76.7	88.2	**77.3**	**70.1**	**59.4**	**52.3**	**48.5**	**46.8**
DDNet	**99.4**	**99.2**	**98.9**	**97.8**	94.7	79.2	25.2	86.1	73.0	60.2	32.5	14.6	7.1	6.0
EvNet	99.2	98.9	98.6	97.6	**95.8**	**90.1**	74.4	**89.8**	71.4	54.5	29.6	18.1	14.4	13.4
	Sensorless							Segment						
PostNet	98.3	19.6	10.9	10.9	11.9	12.4	12.5	98.9	79.6	**57.3**	**31.5**	**18.4**	**20.6**	**19.9**
PriorNet	**99.3**	24.7	11.8	8.6	8.5	8.1	8.3	**99.5**	85.5	40.5	8.9	0.4	0.3	0.2
DDNet	**99.3**	18.0	8.2	6.5	5.4	6.7	7.8	99.2	86.4	36.2	11.9	0.9	0.0	0.0
EvNet	99.0	**42.0**	**28.0**	**17.5**	**13.7**	**13.6**	**14.9**	99.3	**90.6**	55.2	14.2	2.4	0.5	0.1

Table 10. Accuracy under Noise label attacks.

Noise Std	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST							CIFAR10						
PostNet	**99.4**	**98.6**	91.8	**14.9**	**1.3**	**0.1**	0.0	**91.7**	21.5	10.1	0.1	1.2	0.0	1.9
PriorNet	99.3	98.5	**95.7**	14.4	0.0	0.0	0.0	87.7	**28.1**	**11.2**	9.7	5.0	**8.5**	**9.0**
DDNet	**99.4**	**98.6**	92.4	13.3	0.7	0.0	0.0	81.7	23.0	**11.2**	**11.2**	**11.0**	7.8	6.7
EvNet	99.3	96.9	81.6	11.7	0.5	0.0	0.0	89.5	20.7	11.1	5.2	0.5	2.3	3.9
	Sensorless							Segment						
PostNet	98.1	0.1	**3.7**	**11.7**	**11.7**	**11.7**	**11.7**	98.5	39.4	3.9	**1.8**	**12.1**	**20.3**	**22.1**
PriorNet	**99.3**	0.2	0.0	0.0	0.0	0.3	2.4	**99.4**	47.9	8.8	0.0	0.0	0.0	0.0
DDNet	99.0	**0.4**	0.1	0.0	0.0	0.0	0.0	99.1	50.0	**10.3**	0.0	0.0	0.3	0.0
EvNet	98.6	0.2	0.0	0.1	1.4	4.6	8.8	99.1	**50.3**	**10.3**	1.2	0.3	0.0	1.5

6.3.1. UNCERTAINTY ESTIMATION UNDER LABEL ATTACKS

Is low uncertainty a reliable indicator of correct predictions?

On non-perturbed data uncertainty estimates are an indicator of correctly classified samples, but if the input data is perturbed none of the DBU models maintains its high performance. Thus, uncertainty estimates are not a robust indicator of correctly labeled inputs.

Table 2, 11, 12, and 13 illustrate that neither differential entropy nor precision, nor mutual information are a reliable indicator of correct predictions under PGD attacks. DBU-models achieve significantly better results when they are attacked by FGSM-attacks (Table 14), but as FGSM attacks provide much weaker adversarial examples than PGD attacks, this cannot be seen as real advantage.

Table 11. Distinguishing between correctly and wrongly predicted labels based on the differential entropy under PGD label attacks (AUC-PR).

	MNIST							Segment						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
PostNet	99.9	99.9	99.8	98.7	89.5	43.5	9.0	99.9	77.6	31.6	**11.1**	**5.3**	**4.4**	8.7
PriorNet	99.9	99.8	99.6	97.7	90.5	**69.1**	6.4	**100.0**	**96.8**	44.5	4.5	0.4	0.0	**15.2**
DDNet	**100.0**	**100.0**	**99.9**	**99.7**	**97.6**	50.2	0.1	**100.0**	**96.8**	**54.0**	4.3	0.0	0.0	0.0
EvNet	99.6	99.3	98.7	96.1	88.8	63.1	**31.7**	**100.0**	95.9	44.3	5.9	0.8	0.6	0.7

Table 12. Distinguishing between correctly and wrongly predicted labels based on the precision α_0 under PGD label attacks (AUC-PR).

Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST							CIFAR10						
PostNet	**100.0**	99.9	99.7	98.2	87.9	39.1	6.9	**98.7**	88.6	56.2	7.8	1.2	0.4	0.3
PriorNet	99.9	99.8	99.6	97.7	90.4	**69.1**	6.6	92.9	77.7	60.5	**37.6**	**24.9**	**11.3**	**3.0**
DDNet	**100.0**	**100.0**	**100.0**	**99.8**	**98.2**	51.1	0.1	97.6	**91.8**	**78.3**	18.1	0.8	0.0	0.0
EvNet	99.6	99.2	98.6	95.7	88.6	63.6	**32.6**	97.9	85.9	57.2	10.2	4.0	2.4	0.3
	Sensorless							Segment						
PostNet	99.6	7.0	3.3	3.1	**6.9**	**9.8**	**11.3**	99.9	74.2	31.6	**11.1**	**5.0**	**4.2**	**8.6**
PriorNet	99.8	10.5	3.2	0.6	0.2	0.2	1.8	**100.0**	96.9	**45.2**	4.4	0.4	0.0	1.2
DDNet	99.8	8.7	1.3	0.3	0.2	0.1	0.2	**100.0**	**97.1**	45.0	4.1	0.0	0.0	0.0
EvNet	**99.9**	**23.2**	**13.2**	**6.0**	3.7	2.7	2.1	**100.0**	95.7	44.5	5.9	0.8	0.6	0.7

Table 13. Distinguishing between correctly and wrongly predicted labels based on the mutual information under PGD label attacks (AUC-PR).

Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST							CIFAR10						
PostNet	99.7	99.7	99.6	99.2	92.4	40.0	6.9	**97.3**	84.5	56.2	12.2	2.4	0.7	0.3
PriorNet	99.9	99.8	99.6	97.7	90.3	**68.9**	6.4	82.7	65.6	51.4	**35.5**	**24.4**	**11.0**	**2.9**
DDNet	**100.0**	**99.9**	**99.9**	**99.7**	**97.4**	50.2	0.1	96.9	**90.8**	**77.2**	18.8	0.8	0.0	0.0
EvNet	97.8	97.0	95.7	92.6	86.1	62.3	**28.9**	91.3	72.4	47.9	11.4	1.6	0.9	1.6
	Sensorless							Segment						
PostNet	99.3	7.0	3.3	3.3	**7.0**	**9.8**	11.3	99.9	73.2	31.5	**11.1**	**5.0**	**4.3**	**8.7**
PriorNet	**99.8**	10.5	3.2	0.6	0.2	0.1	**11.8**	**100.0**	**96.6**	**45.2**	4.5	0.4	0.0	1.1
DDNet	99.6	8.6	1.3	0.3	0.2	0.1	0.1	**100.0**	96.5	42.4	4.1	0.0	0.0	0.0
EvNet	99.1	**22.0**	**12.6**	**5.9**	3.7	2.7	2.2	**100.0**	90.5	41.0	5.9	0.8	0.6	0.7

Table 14. Distinguishing between correctly and wrongly predicted labels based on the differential entropy under FGSM label attacks (AUC-PR).

Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST							CIFAR10						
PostNet	99.9	99.9	99.8	99.4	97.8	**92.1**	**83.2**	**98.5**	88.7	68.9	31.0	18.6	15.5	16.7
PriorNet	99.9	99.9	99.7	98.3	94.1	88.5	78.6	90.1	73.6	61.6	**46.1**	**38.5**	**35.6**	**37.3**
DDNet	**100.0**	**100.0**	**99.9**	**99.8**	**98.7**	86.4	23.0	97.3	**90.6**	**78.7**	39.4	13.7	6.0	5.1
EvNet	99.6	99.4	99.1	97.8	95.8	90.4	76.8	98.0	86.2	67.4	32.7	19.9	18.2	19.7
	Sensorless							Segment						
PostNet	99.7	11.7	7.3	9.3	11.8	12.5	12.5	99.9	73.6	40.6	**23.7**	**17.2**	**19.8**	**20.2**
PriorNet	99.8	21.4	10.4	8.5	9.0	9.2	10.3	**100.0**	93.7	37.7	5.8	1.1	0.9	0.8
DDNet	99.7	18.5	5.4	4.3	4.2	5.7	7.9	**100.0**	**94.1**	42.9	7.2	1.0	0.0	0.0
EvNet	**99.9**	**44.8**	**29.2**	**18.2**	**15.1**	**14.9**	**15.5**	**100.0**	93.7	**48.7**	8.7	2.4	1.6	0.5

Table 15. Distinguishing between correctly and wrongly predicted labels based on the differential entropy under Noise label attacks (AUC-PR).

Noise Std	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
				MNIST							CIFAR10			
PostNet	99.9	99.8	99.6	**74.2**	**7.4**	**0.2**	0.0	**98.7**	**76.3**	24.3	0.4	4.9	0.0	1.7
PriorNet	99.9	99.9	**99.8**	73.4	0.0	0.0	0.0	85.0	27.8	15.9	**20.4**	7.0	**7.7**	**8.3**
DDNet	**100.0**	**99.9**	99.4	51.1	0.6	0.1	0.0	96.1	61.0	**39.8**	14.2	**11.3**	6.9	6.9
EvNet	99.5	98.4	88.5	20.2	0.9	0.0	0.0	97.5	66.1	21.4	7.7	2.3	3.0	3.8
				Sensorless							Segment			
PostNet	99.7	0.3	**3.2**	**13.3**	**12.0**	**11.7**	**11.7**	99.9	53.9	4.8	1.8	**11.2**	**21.7**	**21.6**
PriorNet	**100.0**	0.3	0.0	0.0	0.0	7.8	11.5	**100.0**	**84.5**	15.6	0.0	0.0	0.0	0.0
DDNet	99.7	**0.9**	0.6	0.0	0.0	0.0	0.0	**100.0**	82.7	**23.9**	0.0	0.0	0.6	0.0
EvNet	99.8	0.3	0.0	0.1	1.7	5.5	10.0	**100.0**	78.3	19.0	**3.5**	0.5	0.0	1.7

Can we use uncertainty estimates to detect attacks against the class prediction?

PGD attacks do not explicitly consider uncertainty during the computation of adversarial examples, but they seem to provide perturbed inputs with similar uncertainty as the original input.

FGSM and Noise attacks are easier to detect, but also weaker thand PGD attacks. This suggests that DBU models are capable of detecting weak attacks by using uncertainty estimation.

Table 16. Attack-Detection based on differential entropy under PGD label attacks (AUC-PR).

	MNIST						Segment					
Att. Rad.	0.1	0.2	0.5	1.0	2.0	4.0	0.1	0.2	0.5	1.0	2.0	4.0
PostNet	57.7	66.3	83.4	90.5	79.0	50.1	**95.6**	73.5	**47.0**	**42.3**	**53.4**	**82.7**
PriorNet	**67.7**	**83.2**	**97.1**	**96.7**	92.1	82.9	86.7	83.3	38.0	31.3	30.8	31.5
DDNet	53.4	57.1	68.5	83.9	**96.0**	**86.3**	76.1	**83.5**	45.4	32.4	30.8	30.8
EvNet	54.8	59.0	68.5	75.9	72.6	59.8	94.9	80.9	41.5	32.5	31.1	31.1

Table 17. Attack-Detection based on precision α_0 under PGD label attacks (AUC-PR).

Att. Rad.	0.1	0.2	0.5	1.0	2.0	4.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST						CIFAR10					
PostNet	63.3	75.7	92.6	95.1	75.3	39.5	**63.4**	**66.9**	42.1	32.9	31.6	31.2
PriorNet	**67.6**	**83.2**	**97.1**	**96.9**	**92.7**	**84.7**	53.3	56.0	55.6	**49.2**	42.2	35.4
DDNet	52.7	55.7	64.7	78.4	91.9	80.9	55.8	60.5	**57.3**	38.7	32.3	31.4
EvNet	49.1	48.0	45.1	42.7	41.8	39.2	48.4	46.9	46.3	46.3	**44.5**	**42.5**
	Sensorless						Segment					
PostNet	39.8	35.8	35.4	**52.0**	**88.2**	**99.0**	**94.6**	70.3	**46.3**	**42.6**	**54.9**	**84.0**
PriorNet	40.9	35.1	32.0	31.1	30.7	30.7	82.7	82.6	39.4	31.6	30.8	30.8
DDNet	**47.7**	**40.3**	35.3	32.8	31.3	30.8	80.0	**86.0**	43.3	33.6	31.0	30.8
EvNet	45.4	39.7	**36.1**	34.8	34.7	36.0	90.9	72.4	40.4	32.4	31.1	31.1

Table 18. Attack-Detection based on mutual information under PGD label attacks (AUC-PR).

Att. Rad.	0.1	0.2	0.5	1.0	2.0	4.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST						CIFAR10					
PostNet	42.2	37.5	36.7	54.5	70.5	70.3	52.2	52.1	50.0	**65.9**	**76.3**	**80.7**
PriorNet	**67.7**	**83.3**	**97.1**	**96.9**	92.6	**84.5**	54.0	56.9	56.3	49.7	42.4	35.5
DDNet	53.1	56.3	66.5	81.0	**94.0**	82.9	**56.0**	**60.8**	**57.4**	38.2	32.1	31.3
EvNet	49.1	48.0	45.2	42.9	41.9	39.3	48.7	47.3	46.3	46.0	44.1	42.2
	Sensorless						Segment					
PostNet	**75.3**	**76.6**	**66.5**	**57.7**	**85.6**	**98.7**	**94.8**	73.5	**55.9**	**47.9**	**58.0**	**84.0**
PriorNet	40.7	35.0	32.0	31.0	30.7	30.7	83.5	82.7	39.2	31.6	30.8	30.8
DDNet	48.0	40.0	35.2	32.6	31.2	30.8	82.4	**88.1**	43.4	33.4	30.9	30.8
EvNet	45.5	39.7	36.1	34.8	34.7	36.0	91.7	72.9	40.5	32.4	31.1	31.1

Table 19. Attack-Detection based on differential entropy under FGSM label attacks (AUC-PR).

Att. Rad.	0.1	0.2	0.5	1.0	2.0	4.0	0.1	0.2	0.5	1.0	2.0	4.0
	MNIST						CIFAR10					
PostNet	55.9	61.8	74.8	84.0	88.9	89.9	**62.1**	**67.2**	65.7	63.1	65.4	73.8
PriorNet	**67.4**	**82.4**	**96.9**	**98.3**	**98.9**	**99.6**	58.4	63.1	68.5	**70.1**	68.5	62.5
DDNet	53.6	57.3	68.3	82.6	95.6	98.7	57.2	62.9	**69.1**	68.7	**69.7**	**76.5**
EvNet	54.1	57.4	63.8	67.6	68.6	69.9	57.8	61.7	63.3	62.9	65.7	72.5
	Sensorless						Segment					
PostNet	**98.4**	**99.8**	**99.9**	**99.9**	**99.9**	**99.9**	**96.9**	**93.9**	**99.5**	**99.9**	**100.0**	**100.0**
PriorNet	48.7	38.6	32.7	32.9	38.6	44.3	89.0	80.8	46.7	37.2	33.7	32.4
DDNet	61.5	47.8	37.1	33.1	32.4	33.2	79.6	86.2	60.2	47.5	36.6	31.6
EvNet	67.3	65.5	72.3	73.4	75.3	79.1	95.7	87.2	59.3	51.7	51.1	53.5

Table 20. Attack-Detection based on differential entropy under Noise label attacks (AUC-PR).

Noise Std.	0.1	0.2	0.5	1.0	2.0	4.0	0.1	0.2	0.5	1.0	2.0	4.0
			MNIST						CIFAR10			
PostNet	51.3	65.3	93.8	95.1	95.2	95.2	**80.8**	**84.5**	**97.6**	**99.5**	99.3	98.2
PriorNet	32.5	36.8	88.9	99.6	99.7	92.7	34.7	32.3	34.3	60.3	95.5	**100.0**
DDNet	**60.7**	**87.6**	**99.8**	**100.0**	**99.9**	**99.8**	59.1	62.6	81.5	98.6	**99.8**	98.7
EvNet	51.2	55.7	66.9	70.3	68.0	67.1	75.7	78.6	88.2	97.8	96.4	95.6
			Sensorless						Segment			
PostNet	**99.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.6**	**99.4**	**100.0**	**100.0**	**100.0**	**100.0**
PriorNet	42.0	33.8	31.5	34.7	43.7	47.0	56.7	56.7	39.8	33.7	31.9	33.7
DDNet	53.4	43.5	34.3	31.6	32.5	36.1	57.0	58.9	43.1	33.7	31.5	31.3
EvNet	67.1	78.8	88.3	95.4	96.9	97.8	60.8	63.5	61.2	64.8	73.7	85.2

6.3.2. Attacking uncertainty estimation

Are uncertainty estimates a robust feature for OOD detection?

Using uncertainty estimation to distinguish between ID and OOD data is not robust as shown in the following tables.

Table 21. OOD detection based on differential entropy under PGD uncertainty attacks against differential entropy on ID data and OOD data (AUC-PR).

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
MNIST – KMNIST														
PostNet	94.5	94.1	93.9	91.1	77.1	44.0	31.9	94.5	93.1	91.4	82.1	62.2	50.7	48.8
PriorNet	**99.6**	**99.4**	**99.1**	**97.8**	**93.8**	**77.6**	**32.0**	**99.6**	**99.4**	**99.1**	98.0	94.6	85.5	**73.9**
DDNet	99.3	99.1	98.9	**97.8**	93.5	63.3	30.7	99.3	99.1	99.0	**98.3**	**96.7**	**91.3**	73.8
EvNet	69.0	67.1	65.6	61.8	57.4	50.9	43.6	69.0	55.8	48.0	39.4	36.2	34.9	34.4
Seg. – Seg. class sky														
PostNet	**99.0**	**80.7**	**53.5**	**38.0**	**34.0**	**41.6**	**49.5**	**99.0**	**88.4**	69.2	45.1	**36.4**	**42.6**	**75.4**
PriorNet	34.8	31.4	30.9	30.8	30.8	30.8	30.8	34.8	31.8	31.0	30.8	30.8	30.8	32.1
DDNet	31.5	30.9	30.8	30.8	30.8	30.8	30.8	31.5	31.0	30.8	30.8	30.8	30.8	30.8
EvNet	92.5	67.2	43.2	31.6	30.9	30.9	31.2	92.5	86.1	**82.7**	**48.9**	32.7	30.9	30.9

Table 22. OOD detection under PGD uncertainty attacks against differential entropy on ID data and OOD data (AUC-ROC).

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
MNIST – KMNIST														
PostNet	91.6	91.3	91.9	91.5	80.2	38.8	9.2	91.6	90.4	89.0	81.6	62.6	45.0	43.1
PriorNet	**99.8**	**99.7**	**99.5**	**99.0**	**97.1**	**81.1**	8.7	**99.8**	**99.7**	**99.6**	**99.1**	**97.7**	**93.0**	**84.9**
DDNet	99.2	98.9	98.6	97.3	92.1	58.2	1.2	99.2	99.0	98.8	97.9	95.8	89.1	69.3
EvNet	81.2	79.6	78.2	74.6	69.5	58.7	**43.0**	81.2	67.2	54.8	35.4	25.5	20.7	18.5
CIFAR10 – SVHN														
PostNet	87.0	71.9	56.3	**30.2**	**20.2**	**15.0**	**9.7**	87.0	71.0	54.3	33.5	30.3	26.2	19.4
PriorNet	62.4	48.2	35.9	13.8	3.6	0.9	0.3	62.4	48.0	35.6	14.8	6.6	3.4	1.6
DDNet	87.0	**76.0**	**63.6**	29.3	6.1	1.1	0.4	87.0	**78.1**	**66.1**	26.2	5.1	0.7	0.1
EvNet	**88.0**	69.1	51.7	24.6	15.5	9.5	4.2	**88.0**	72.0	60.7	**47.9**	**42.1**	**33.3**	**24.0**
Sens. – Sens. class 10, 11														
PostNet	**85.3**	**49.1**	**38.1**	**7.8**	**8.2**	8.2	8.2	**85.3**	**57.2**	**54.0**	**27.3**	**31.5**	**86.7**	**99.5**
PriorNet	28.1	0.8	0.3	0.4	1.6	**8.4**	**26.8**	28.1	2.5	0.7	0.2	2.3	18.9	41.0
DDNet	21.0	3.0	0.9	0.4	0.6	2.1	7.3	21.0	4.4	2.1	1.9	2.2	2.2	4.1
EvNet	74.2	21.4	12.2	4.3	1.4	0.6	0.3	74.2	45.3	38.5	19.6	9.6	12.1	26.0
Seg. – Seg. class sky														
PostNet	**99.2**	**84.7**	**55.5**	**23.0**	**9.7**	**4.4**	**4.7**	**99.2**	**92.1**	**77.1**	41.5	**24.9**	**41.0**	**80.8**
PriorNet	17.1	4.4	1.3	0.0	0.0	0.0	0.1	17.1	5.9	1.5	0.1	0.0	0.1	5.8
DDNet	4.1	1.1	0.0	0.0	0.0	0.0	0.0	4.1	1.8	0.4	0.0	0.0	0.0	0.0
EvNet	91.2	54.5	23.3	3.9	0.9	0.4	0.2	91.2	82.9	76.4	**42.2**	9.7	0.8	0.6

Table 23. OOD detection (AU-PR) under PGD uncertainty attacks against precision α_0 on ID data and OOD data.

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
							MNIST – KMNIST							
PostNet	98.4	97.4	96.0	88.8	70.9	39.3	31.3	98.4	97.2	95.2	82.8	52.6	34.3	32.1
PriorNet	**99.6**	**99.5**	**99.2**	**98.0**	**94.1**	**76.0**	31.1	**99.6**	**99.5**	**99.2**	**98.2**	**95.3**	**87.5**	**75.6**
DDNet	97.2	96.7	96.1	93.8	86.4	53.2	31.0	97.2	96.7	96.2	94.5	91.1	82.9	64.6
EvNet	39.8	39.2	38.8	37.9	37.1	36.3	**35.4**	39.8	34.5	32.5	31.2	31.0	30.9	31.0
							CIFAR10 – SVHN							
PostNet	**82.4**	63.8	46.1	22.3	17.4	16.7	16.4	**82.4**	61.8	41.5	21.8	**19.8**	**17.5**	**15.8**
PriorNet	37.9	25.0	19.2	15.8	15.4	15.4	15.4	37.9	25.9	19.4	15.6	15.4	15.4	15.4
DDNet	81.1	**70.1**	**58.4**	**30.0**	16.7	15.5	15.4	81.1	**71.2**	**59.9**	**27.8**	16.5	15.5	15.4
EvNet	34.7	27.4	25.4	22.0	**19.7**	**18.1**	**17.1**	34.7	19.4	18.1	17.1	16.8	16.2	15.7
							Sens. – Sens. class 10, 11							
PostNet	**77.4**	**39.6**	**35.9**	**31.7**	**44.4**	**44.4**	**44.4**	**77.4**	40.3	**38.6**	29.5	**34.0**	**79.4**	**97.4**
PriorNet	35.9	27.0	26.8	26.8	26.8	27.5	36.2	35.9	27.7	27.0	26.7	26.6	26.5	26.5
DDNet	55.6	34.4	31.7	30.4	29.5	30.2	33.4	55.6	**40.9**	34.1	28.0	26.9	26.6	26.5
EvNet	66.3	33.3	29.7	27.0	27.1	29.2	33.9	66.3	39.3	37.1	**31.3**	28.3	28.4	29.7
							Seg. – Seg. class sky							
PostNet	**98.4**	74.8	51.0	**37.2**	**32.8**	**43.5**	**49.9**	**98.4**	84.7	66.1	42.4	34.8	**40.9**	**71.2**
PriorNet	32.1	30.9	30.8	30.8	30.8	30.8	30.8	32.1	31.0	30.8	30.8	30.8	30.8	30.8
DDNet	31.0	30.8	30.8	30.8	30.8	30.8	30.8	31.0	30.8	30.8	30.8	30.8	30.8	30.8
EvNet	98.3	**83.0**	**60.5**	34.0	31.0	30.8	30.8	98.3	**94.4**	**88.8**	**65.6**	**37.0**	31.4	30.9

Table 24. OOD detection (AUC-ROC) under PGD uncertainty attacks against precision α_0 on ID data and OOD data.

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
							MNIST – KMNIST							
PostNet	98.4	97.6	96.4	90.9	74.0	28.9	6.3	98.4	97.6	96.3	89.0	61.3	19.6	9.7
PriorNet	**99.8**	**99.7**	**99.6**	**99.1**	**97.2**	**79.4**	4.4	**99.8**	**99.7**	**99.6**	**99.2**	**98.0**	**93.9**	**85.8**
DDNet	96.5	95.9	95.1	92.0	82.6	44.3	3.5	96.5	95.9	95.2	92.9	88.6	78.7	59.4
EvNet	35.9	34.1	32.8	30.1	27.4	24.6	**21.4**	35.9	18.7	10.4	3.7	2.0	1.7	2.0
							CIFAR10 – SVHN							
PostNet	**87.4**	71.2	54.8	29.2	19.0	14.0	9.4	**87.4**	71.4	54.1	30.1	**25.8**	**17.5**	**5.8**
PriorNet	45.6	31.1	20.4	6.3	1.4	0.3	0.1	45.6	32.2	21.7	5.4	1.0	0.3	0.1
DDNet	84.9	**73.8**	**61.8**	30.2	9.3	3.0	0.8	84.9	**76.6**	**66.2**	**34.6**	10.4	2.3	0.3
EvNet	61.2	49.4	45.2	**37.6**	**30.5**	**23.4**	**17.0**	61.2	29.4	23.0	16.8	14.2	10.2	5.5
							Sens. – Sens. class 10, 11							
PostNet	**87.2**	**48.8**	**37.3**	4.1	0.7	0.7	0.7	**87.2**	**50.0**	**45.4**	16.5	**27.6**	**81.9**	**98.0**
PriorNet	37.3	3.5	2.4	2.2	2.9	6.3	**19.2**	37.3	8.0	3.6	1.4	0.6	0.1	0.0
DDNet	55.2	23.7	17.7	**14.1**	**12.5**	**12.7**	15.7	55.2	37.1	27.7	9.4	2.5	0.6	0.1
EvNet	75.5	30.8	18.2	5.8	1.6	0.6	0.2	75.5	47.8	41.9	**24.1**	10.2	10.2	15.6
							Seg. – Seg. class sky							
PostNet	**98.6**	77.7	**50.8**	**20.3**	**8.2**	**1.3**	**0.5**	**98.6**	88.9	73.4	36.2	19.4	**36.7**	**75.2**
PriorNet	8.5	1.3	0.2	0.0	0.0	0.0	0.1	8.5	2.0	0.4	0.0	0.0	0.0	0.0
DDNet	2.2	0.3	0.0	0.0	0.0	0.0	0.0	2.2	0.5	0.1	0.0	0.0	0.0	0.0
EvNet	97.7	**78.4**	47.7	9.9	1.2	0.2	0.1	97.7	**93.5**	**86.9**	**62.2**	**21.5**	3.7	1.0

Table 25. OOD detection (AU-PR) under PGD uncertainty attacks against distributional uncertainty on ID data and OOD data.

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
MNIST – KMNIST														
PostNet	80.5	76.2	73.4	69.1	66.6	65.4	**60.2**	80.5	72.1	63.9	43.9	33.0	30.9	30.8
PriorNet	**99.6**	**99.4**	**99.2**	**98.0**	**94.1**	**76.3**	31.2	**99.6**	**99.4**	**99.2**	**98.2**	**95.2**	**87.2**	**75.2**
DDNet	98.4	98.1	97.7	95.8	89.5	56.2	30.9	98.4	98.1	97.8	96.5	93.8	86.3	67.7
EvNet	40.1	39.5	39.1	38.2	37.3	36.5	35.6	40.1	34.6	32.6	31.3	31.0	31.0	31.1
CIFAR10 – SVHN														
PostNet	64.2	44.7	37.5	**31.1**	**28.5**	**25.0**	**19.3**	64.2	31.0	19.5	16.3	16.4	**16.5**	**16.3**
PriorNet	40.8	27.4	20.4	15.9	15.4	15.4	15.4	40.8	28.3	21.1	15.9	15.4	15.4	15.4
DDNet	**82.0**	**71.0**	**59.1**	29.9	16.6	15.5	15.4	**82.0**	**72.2**	**60.3**	**26.3**	16.2	15.4	15.4
EvNet	36.4	28.7	26.5	22.8	20.2	18.4	17.2	36.4	19.8	18.3	17.2	**16.9**	16.2	15.7
Sens. – Sens. class 10, 11														
PostNet	**79.1**	**40.3**	**35.9**	**33.0**	**45.5**	**45.5**	45.5	**79.1**	**47.3**	**43.7**	**36.5**	**37.9**	**74.6**	**96.5**
PriorNet	35.5	26.8	26.7	26.9	29.6	43.7	**68.7**	35.5	27.5	26.9	26.7	26.6	26.5	26.5
DDNet	52.9	31.7	29.8	29.1	28.4	30.1	37.6	52.9	38.4	31.5	27.5	26.8	26.6	26.5
EvNet	66.3	33.3	29.6	27.0	27.2	29.3	35.2	66.3	39.3	37.1	31.3	28.3	28.4	29.7
Seg. – Seg. class sky														
PostNet	98.0	76.3	53.1	**37.4**	**32.9**	**44.6**	**50.2**	98.0	83.5	64.8	41.8	35.4	**43.1**	**71.3**
PriorNet	32.3	30.9	30.8	30.8	30.8	32.5	45.0	32.3	31.0	30.8	30.8	30.8	30.8	30.8
DDNet	30.9	30.8	30.8	30.8	30.8	30.8	30.8	30.9	30.8	30.8	30.8	30.8	30.8	30.8
EvNet	**98.1**	**82.1**	**59.1**	33.8	31.0	30.8	30.8	**98.1**	**93.8**	**88.2**	**64.5**	**36.4**	31.3	31.0

Table 26. OOD detection (AUC-ROC) under PGD uncertainty attacks against distributional uncertainty on ID data and OOD data.

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
MNIST – KMNIST														
PostNet	90.1	88.0	86.2	82.2	79.0	77.1	**66.1**	90.1	84.5	77.2	46.4	12.9	2.7	2.4
PriorNet	**99.8**	**99.7**	**99.6**	**99.1**	**97.2**	**79.7**	4.7	**99.8**	**99.7**	**99.6**	**99.2**	**97.9**	**93.7**	**85.6**
DDNet	98.1	97.7	97.2	94.8	87.0	48.7	3.0	98.1	97.8	97.3	95.8	92.3	83.3	63.3
EvNet	36.8	35.0	33.7	30.9	28.2	25.3	22.1	36.8	19.3	10.7	3.9	2.1	1.8	2.2
CIFAR10 – SVHN														
PostNet	82.9	67.7	59.2	**51.3**	**47.7**	**40.1**	**24.2**	82.9	51.9	26.2	8.9	9.5	**11.1**	**9.9**
PriorNet	48.0	33.6	22.5	7.1	1.6	0.3	0.1	48.0	34.8	24.0	6.7	1.6	0.6	0.2
DDNet	**85.9**	**74.9**	**62.7**	30.1	8.3	2.3	0.6	**85.9**	**77.6**	**66.9**	**32.1**	8.0	1.5	0.2
EvNet	63.3	51.4	47.1	39.3	32.1	24.9	17.9	63.3	31.1	24.4	17.7	**15.0**	10.7	5.7
Sens. – Sens. class 10, 11														
PostNet	**87.1**	**50.9**	**37.8**	5.5	4.5	4.5	4.5	**87.1**	**55.3**	**51.1**	**34.4**	**38.9**	**79.7**	**97.9**
PriorNet	36.5	2.9	1.8	1.8	5.2	**21.5**	**52.8**	36.5	7.3	3.0	1.3	0.5	0.1	0.0
DDNet	52.3	18.7	13.1	**10.3**	**9.3**	10.8	18.4	52.3	33.1	22.0	6.7	2.2	0.6	0.1
EvNet	75.5	30.7	18.1	5.8	1.6	0.6	0.8	75.5	47.7	41.8	23.8	10.3	10.2	15.8
Seg. – Seg. class sky														
PostNet	**98.6**	**78.3**	**51.9**	**20.5**	**8.3**	**2.1**	1.7	**98.6**	88.8	73.1	35.9	**21.4**	**39.9**	**75.9**
PriorNet	9.4	1.6	0.3	0.0	0.0	1.8	**15.4**	9.4	2.4	0.4	0.0	0.0	0.0	0.0
DDNet	1.3	0.2	0.0	0.0	0.0	0.0	0.0	1.3	0.2	0.0	0.0	0.0	0.0	0.0
EvNet	97.4	77.1	45.9	9.4	1.3	0.2	0.1	97.4	**92.9**	**86.1**	**60.9**	20.4	3.0	1.2

Table 27. OOD detection (AU-PR) under FGSM uncertainty attacks against differential entropy on ID data and OOD data.

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
							MNIST – KMNIST							
PostNet	94.5	94.2	94.1	93.5	89.9	81.2	**71.6**	94.5	93.3	92.0	87.6	81.1	75.7	75.7
PriorNet	**99.6**	**99.4**	**99.2**	**98.1**	**95.6**	**90.0**	65.3	**99.6**	**99.4**	**99.2**	**98.6**	97.5	**95.9**	**94.4**
DDNet	99.3	99.1	98.9	98.0	95.4	80.9	48.2	99.3	99.2	99.0	98.5	**97.6**	95.5	92.0
EvNet	69.0	67.4	66.2	64.0	61.9	59.8	56.70	9.0	60.1	56.5	53.4	52.7	52.9	53.5
							CIFAR10 – SVHN							
PostNet	81.8	66.2	61.6	**64.2**	**65.7**	61.3	48.4	81.8	63.1	51.9	43.4	46.6	**61.7**	**77.0**
PriorNet	54.4	40.6	33.8	27.0	25.5	27.2	35.5	54.4	42.3	36.8	30.6	28.3	29.5	32.1
DDNet	**82.8**	**71.9**	**64.6**	53.8	50.2	47.8	41.0	**82.8**	**71.5**	**60.5**	39.1	31.4	41.2	66.6
EvNet	80.3	67.8	64.0	61.9	61.6	57.4	**49.6**	80.3	59.2	51.5	**46.7**	**49.0**	56.3	64.6
							Sens. – Sens. class 10, 11							
PostNet	**74.5**	40.6	37.2	31.4	38.1	44.9	45.9	**74.5**	**99.6**	**99.8**	**99.9**	**99.9**	**99.9**	**99.9**
PriorNet	32.3	35.7	**57.6**	**83.1**	**88.8**	79.7	70.0	32.3	28.3	28.1	27.6	28.0	32.7	38.5
DDNet	31.7	31.3	44.4	70.3	87.9	**92.5**	**91.9**	31.7	28.8	29.3	29.1	27.7	27.9	28.01
EvNet	66.5	**45.7**	46.8	42.3	42.0	41.4	41.8	66.5	54.7	66.5	76.2	71.1	75.3	75.8
							Seg. – Seg. class sky							
PostNet	**99.0**	**80.8**	**66.4**	43.6	37.0	35.5	43.0	**99.0**	**94.8**	**92.0**	**98.5**	**99.7**	**100.0**	**100.0**
PriorNet	34.8	31.2	31.4	46.3	**74.0**	**88.8**	**94.5**	34.8	31.6	31.0	31.2	30.9	30.8	30.8
DDNet	31.5	30.8	30.8	30.9	37.9	56.2	84.3	31.5	30.9	30.8	30.8	30.8	30.8	30.8
EvNet	92.5	64.9	54.6	**66.6**	69.5	69.6	64.6	92.5	85.9	83.0	66.3	66.1	61.1	56.8

Table 28. OOD detection (AU-PR) under Noise uncertainty attacks against differential entropy on ID data and OOD data.

	ID-Attack (non-attacked OOD)							OOD-Attack (non-attacked ID)						
Noise Std	0.0	0.1	0.2	0.5	1.0	2.0	4.0	0.0	0.1	0.2	0.5	1.0	2.0	4.0
							MNIST – KMNIST							
PostNet	93.0	94.2	82.3	34.4	31.6	31.0	30.9	92.2	91.8	91.5	92.3	92.7	93.2	93.5
PriorNet	**99.7**	**99.6**	**96.7**	**40.0**	**40.6**	**45.7**	**55.6**	**99.5**	97.3	96.5	99.4	**100.0**	99.5	72.4
DDNet	99.1	97.5	81.2	31.3	31.0	30.9	31.2	99.0	**98.8**	**99.2**	**99.8**	99.9	**99.8**	**99.1**
EvNet	65.5	60.5	51.4	35.3	34.5	35.5	35.0	62.5	47.2	40.9	35.1	34.6	33.5	34.9
							CIFAR10 – SVHN							
PostNet	88.5	41.4	39.8	31.0	30.7	31.6	33.9	88.5	**86.6**	**81.9**	**93.0**	**98.5**	98.6	97.3
PriorNet	73.3	88.3	**95.3**	**92.4**	**70.4**	30.9	30.8	73.3	31.6	30.9	31.7	51.8	94.3	**100.0**
DDNet	87.3	69.3	78.4	55.2	31.6	30.7	31.4	87.3	55.8	57.9	73.9	97.3	**99.5**	97.2
EvNet	**92.4**	**56.8**	53.8	33.4	30.9	**32.9**	**36.6**	**92.4**	73.7	73.5	77.7	93.7	92.5	92.1
							Sens. – Sens. class 10, 11							
PostNet	**85.3**	**30.8**	**39.4**	50.0	50.0	50.0	50.0	**85.3**	**98.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
PriorNet	32.3	**30.8**	34.9	**83.7**	**77.7**	49.8	**80.3**	32.3	30.7	30.7	32.5	40.1	49.9	47.6
DDNet	31.1	30.7	30.7	32.4	58.8	**88.1**	74.3	31.1	30.7	30.7	30.7	30.8	31.6	39.1
EvNet	80.3	**30.8**	31.2	37.9	46.3	50.0	50.0	80.3	34.6	38.4	53.9	69.3	78.8	81.5
							Seg. – Seg. class sky							
PostNet	**99.9**	**41.8**	30.8	**34.5**	**49.1**	50.0	50.0	**99.9**	**97.4**	**96.6**	**99.5**	**100.0**	**100.0**	**100.0**
PriorNet	31.0	30.8	30.8	30.8	32.7	**69.0**	78.3	31.0	30.8	30.8	30.8	30.9	31.1	32.4
DDNet	30.8	30.8	30.8	30.8	30.8	58.2	**91.3**	30.8	30.8	30.8	30.8	30.8	30.8	31.9
EvNet	99.1	38.1	**32.2**	30.8	30.8	32.2	37.5	99.1	95.6	87.6	58.0	44.9	46.6	53.8

6.4. How to make DBU models more robust

To improve robustness of DBU models we perform median smoothing and adversarial training. Smoothing computes the smooth median, worst case and best case performance of DBU models for three tasks: distinguishing between correct and wrong predictions, attack detection, distinguishing between ID data and OOD data under label attacks and under uncertainty attacks.

Table 29. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under PGD label attacks. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	80.5 · **91.5** · 94.5	52.8 · **71.6** · 95.2	31.9 · **51.0** · 96.8	5.6 · **11.7** · 100.0	0.3 · **0.6** · 100.0	0.0 · **0.0** · 100.0
	PriorNet	81.9 · **86.8** · 88.0	69.6 · **78.0** · 90.1	50.9 · **65.8** · 89.4	36.5 · **59.9** · 97.0	24.3 · **39.3** · 100.0	9.2 · **17.9** · 100.0
	DDNet	65.9 · **81.2** · 83.0	55.8 · **70.5** · 87.2	37.8 · **56.8** · 88.1	10.1 · **21.9** · 94.3	0.9 · **1.6** · 99.6	0.0 · **0.0** · 100.0
	EvNet	76.3 · **90.2** · 91.7	54.7 · **74.3** · 95.7	31.6 · **51.5** · 94.5	5.8 · **11.9** · 86.9	1.9 · **7.0** · 100.0	1.1 · **4.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	52.1 · **71.8** · 95.6	31.2 · **47.9** · 96.1	7.8 · **14.7** · 98.6	1.8 · **4.4** · 100.0	0.3 · **0.5** · 100.0
	PriorNet	-	57.6 · **71.7** · 88.9	46.1 · **64.5** · 90.1	38.1 · **59.3** · 99.5	32.3 · **51.7** · 100.0	22.1 · **41.6** · 97.4
	DDNet	-	58.6 · **78.4** · 92.2	49.4 · **66.0** · 90.5	12.0 · **21.4** · 98.1	0.8 · **1.0** · 96.6	0.0 · **0.0** · 100.0
	EvNet	-	24.3 · **34.2** · 51.8	32.6 · **49.5** · 95.5	5.9 · **13.0** · 100.0	2.6 · **5.2** · 99.9	2.9 · **5.9** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	52.8 · **74.2** · 94.6	33.0 · **49.4** · 87.5	7.7 · **14.2** · 99.0	0.6 · **1.2** · 100.0	0.7 · **1.1** · 100.0
	PriorNet	-	50.6 · **68.1** · 88.6	44.4 · **66.1** · 96.0	35.1 · **57.4** · 98.4	18.4 · **32.2** · 100.0	15.2 · **29.3** · 100.0
	DDNet	-	68.8 · **84.4** · 93.2	45.1 · **60.8** · 86.8	12.3 · **22.0** · 91.0	0.8 · **1.7** · 87.0	0.0 · **0.0** · 100.0
	EvNet	-	54.2 · **73.7** · 96.1	30.5 · **50.0** · 99.5	7.1 · **13.9** · 100.0	3.7 · **8.7** · 75.2	3.3 · **5.8** · 100.0

Table 30. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under PGD label attacks. Smoothed DBU models on MNIST. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	97.2 · **99.4** · 100.0	95.9 · **99.1** · 99.9	94.7 · **98.9** · 99.9	89.3 · **96.8** · 99.9	75.5 · **90.2** · 100.0	35.5 · **56.7** · 100.0
	PriorNet	96.8 · **99.2** · 99.3	95.5 · **99.1** · 99.7	94.6 · **98.8** · 99.7	90.2 · **97.2** · 99.9	81.1 · **93.4** · 99.9	53.9 · **75.2** · 100.0
	DDNet	97.6 · **99.4** · 99.5	96.8 · **99.2** · 99.4	95.5 · **98.8** · 99.4	90.4 · **97.2** · 99.8	77.0 · **91.3** · 100.0	29.2 · **48.6** · 100.0
	EvNet	97.3 · **99.4** · 99.4	95.4 · **98.8** · 99.6	93.9 · **98.7** · 99.9	89.0 · **96.5** · 100.0	78.9 · **92.9** · 100.0	52.2 · **73.2** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	94.4 · **98.6** · 99.5	90.6 · **97.9** · 99.9	83.4 · **93.1** · 99.9	72.1 · **91.2** · 100.0	41.8 · **65.0** · 100.0
	PriorNet	-	94.4 · **98.5** · 99.5	93.6 · **98.8** · 99.8	89.1 · **96.6** · 99.8	81.5 · **94.5** · 100.0	71.6 · **88.4** · 100.0
	DDNet	-	94.9 · **98.3** · 98.7	94.6 · **97.9** · 98.9	88.2 · **97.4** · 99.8	72.1 · **89.3** · 100.0	28.1 · **49.3** · 100.0
	EvNet	-	88.8 · **95.3** · 97.9	91.5 · **97.1** · 99.4	85.2 · **94.9** · 100.0	78.1 · **91.4** · 100.0	54.3 · **75.3** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	92.8 · **98.3** · 99.8	92.5 · **98.3** · 99.9	86.2 · **94.8** · 99.8	71.0 · **89.5** · 100.0	34.6 · **54.2** · 100.0
	PriorNet	-	95.1 · **98.6** · 99.6	94.1 · **98.0** · 99.4	87.7 · **97.2** · 99.9	80.2 · **93.4** · 100.0	68.5 · **87.8** · 100.0
	DDNet	-	96.0 · **98.4** · 98.8	95.0 · **97.6** · 98.7	87.6 · **95.3** · 99.7	73.9 · **90.2** · 100.0	32.8 · **54.4** · 100.0
	EvNet	-	93.3 · **98.6** · 99.5	89.8 · **97.2** · 99.2	86.2 · **95.4** · 100.0	82.1 · **93.7** · 100.0	52.4 · **73.3** · 100.0

Table 31. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under PGD label attacks. Smoothed DBU models on Sensorless. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	93.5 · **98.4** · 100.0	6.7 · **12.4** · 100.0	2.9 · **5.3** · 100.0	4.1 · **4.1** · 49.1	6.4 · **6.4** · 6.4	10.6 · **10.6** · 10.6
	PriorNet	97.1 · **99.3** · 100.0	8.6 · **17.6** · 100.0	3.3 · **7.7** · 100.0	0.7 · **1.5** · 100.0	0.4 · **0.7** · 100.0	0.1 · **0.2** · 100.0
	DDNet	95.9 · **98.9** · 99.7	7.0 · **14.0** · 100.0	0.8 · **1.3** · 100.0	0.2 · **0.4** · 100.0	0.2 · **0.2** · 100.0	0.2 · **0.4** · 100.0
	EvNet	94.0 · **99.0** · 99.9	18.1 · **34.2** · 100.0	9.6 · **17.1** · 100.0	4.1 · **6.8** · 100.0	2.7 · **4.9** · 100.0	2.4 · **4.3** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	7.9 · **14.9** · 100.0	2.9 · **6.3** · 100.0	6.6 · **6.6** · 6.6	7.2 · **7.2** · 7.2	9.6 · **9.6** · 9.6
	PriorNet	-	18.1 · **32.1** · 100.0	8.7 · **16.7** · 100.0	0.1 · **0.2** · 100.0	0.0 · **0.0** · 100.0	0.8 · **1.0** · 100.0
	DDNet	-	6.9 · **13.4** · 100.0	4.3 · **9.0** · 100.0	0.2 · **0.3** · 100.0	0.2 · **0.4** · 100.0	0.2 · **0.8** · 100.0
	EvNet	-	19.7 · **35.7** · 100.0	9.4 · **16.2** · 100.0	1.6 · **3.0** · 100.0	2.5 · **5.6** · 100.0	1.0 · **1.8** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	7.9 · **14.4** · 100.0	4.8 · **9.3** · 100.0	6.6 · **6.6** · 6.6	6.7 · **6.7** · 6.7	10.6 · **10.6** · 10.6
	PriorNet	-	19.1 · **32.7** · 100.0	6.9 · **13.7** · 100.0	0.7 · **1.7** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	-	5.4 · **10.2** · 100.0	0.7 · **1.8** · 100.0	0.5 · **0.9** · 100.0	0.3 · **1.2** · 100.0	0.2 · **0.6** · 100.0
	EvNet	-	22.3 · **38.4** · 100.0	11.7 · **22.4** · 100.0	7.1 · **13.1** · 100.0	1.8 · **3.4** · 100.0	0.6 · **1.0** · 100.0

Table 32. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under PGD label attacks. Smoothed DBU models on Segment. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model)..

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	94.0 · **99.1** · 99.8	63.5 · **84.7** · 100.0	33.2 · **56.1** · 100.0	10.2 · **16.9** · 100.0	5.2 · **10.3** · 100.0	0.3 · **0.3** · 0.3
	PriorNet	97.0 · **99.8** · 99.9	75.6 · **90.8** · 100.0	31.1 · **50.8** · 100.0	2.6 · **4.7** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	96.2 · **99.5** · 99.7	75.7 · **89.8** · 99.9	28.5 · **51.6** · 100.0	3.7 · **8.2** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	EvNet	95.8 · **99.6** · 99.9	80.2 · **93.7** · 100.0	35.2 · **57.2** · 100.0	6.8 · **12.0** · 100.0	1.2 · **2.1** · 100.0	1.1 · **2.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	66.0 · **85.5** · 100.0	22.5 · **41.0** · 100.0	9.0 · **16.3** · 100.0	5.2 · **9.7** · 100.0	0.6 · **0.6** · 0.6
	PriorNet	-	79.0 · **92.4** · 100.0	45.2 · **68.8** · 100.0	9.2 · **13.9** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	-	76.2 · **91.0** · 99.6	27.2 · **45.3** · 100.0	2.3 · **4.3** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	EvNet	-	82.7 · **95.2** · 100.0	34.0 · **53.8** · 100.0	10.9 · **23.2** · 100.0	0.5 · **4.2** · 100.0	2.1 · **5.1** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	71.5 · **87.6** · 100.0	33.5 · **54.5** · 100.0	12.8 · **25.6** · 100.0	6.5 · **10.3** · 87.2	0.0 · **0.0** · 100.0
	PriorNet	-	82.1 · **96.5** · 100.0	44.1 · **65.4** · 100.0	9.0 · **15.7** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	-	77.4 · **91.4** · 99.9	29.4 · **50.3** · 100.0	4.0 · **6.5** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	EvNet	-	76.2 · **90.7** · 100.0	35.7 · **55.4** · 100.0	4.2 · **6.4** · 100.0	0.8 · **1.4** · 100.0	0.0 · **0.0** · 100.0

Table 33. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under FGSM label attacks. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	80.5 · **91.4** · 94.4	52.3 · **73.2** · 95.4	35.8 · **57.2** · 97.5	17.0 · **29.0** · 100.0	10.2 · **18.7** · 100.0	8.1 · **14.7** · 100.0
	PriorNet	81.9 · **87.7** · 88.8	69.6 · **78.4** · 90.3	53.3 · **70.5** · 91.7	42.1 · **62.6** · 97.2	37.5 · **55.7** · 100.0	36.0 · **59.5** · 100.0
	DDNet	65.9 · **84.1** · 85.6	55.3 · **69.6** · 87.0	38.6 · **55.8** · 87.2	16.3 · **28.5** · 94.6	6.4 · **12.0** · 99.9	3.6 · **7.2** · 100.0
	EvNet	76.3 · **90.4** · 91.7	54.1 · **74.5** · 95.5	35.5 · **54.7** · 95.1	14.6 · **29.3** · 95.6	8.6 · **16.1** · 100.0	7.2 · **13.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	52.3 · **71.6** · 95.1	34.7 · **54.8** · 96.6	18.9 · **32.1** · 99.4	10.9 · **19.2** · 100.0	8.5 · **16.2** · 100.0
	PriorNet	-	58.1 · **69.6** · 87.6	47.1 · **65.7** · 90.3	40.2 · **59.5** · 99.3	36.2 · **59.5** · 100.0	25.1 · **42.1** · 97.7
	DDNet	-	57.1 · **75.2** · 91.0	49.3 · **65.3** · 90.5	18.4 · **33.6** · 98.5	7.6 · **13.5** · 99.9	3.3 · **9.6** · 100.0
	EvNet	-	24.1 · **36.5** · 54.2	37.1 · **56.7** · 96.7	16.2 · **29.9** · 100.0	11.4 · **21.8** · 100.0	13.0 · **26.1** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	52.0 · **71.8** · 94.5	35.8 · **54.6** · 89.9	18.4 · **33.6** · 99.8	10.2 · **19.1** · 100.0	12.2 · **23.0** · 100.0
	PriorNet	-	50.6 · **67.3** · 88.5	46.2 · **64.3** · 95.1	39.9 · **60.8** · 98.5	27.7 · **46.2** · 100.0	28.5 · **48.6** · 100.0
	DDNet	-	67.7 · **82.2** · 92.4	45.7 · **64.7** · 88.8	20.5 · **34.8** · 93.6	6.1 · **13.1** · 91.8	4.1 · **8.4** · 100.0
	EvNet	-	53.9 · **73.6** · 96.3	34.2 · **55.3** · 99.7	16.1 · **31.2** · 100.0	6.1 · **13.5** · 86.1	18.1 · **34.0** · 100.0

Table 34. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under FGSM label attacks. Smoothed DBU models on MNIST. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	97.2 · **99.3** · 99.9	96.1 · **99.2** · 99.9	95.2 · **98.9** · 99.9	91.7 · **98.0** · 99.9	86.1 · **95.9** · 100.0	75.7 · **91.1** · 100.0
	PriorNet	96.8 · **99.2** · 99.3	95.5 · **99.0** · 99.6	94.7 · **98.7** · 99.6	91.3 · **97.6** · 99.9	85.5 · **95.6** · 100.0	78.7 · **92.4** · 100.0
	DDNet	97.6 · **99.3** · 99.4	96.8 · **99.2** · 99.5	95.6 · **98.7** · 99.4	91.7 · **97.7** · 99.9	83.4 · **95.2** · 100.0	58.3 · **79.6** · 100.0
	EvNet	97.3 · **99.3** · 99.4	95.5 · **99.0** · 99.6	94.3 · **98.9** · 99.9	92.0 · **97.7** · 100.0	87.4 · **96.3** · 100.0	78.8 · **92.4** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	95.1 · **98.9** · 99.8	91.2 · **97.2** · 99.6	87.6 · **96.3** · 99.9	81.0 · **93.3** · 100.0	69.9 · **87.2** · 100.0
	PriorNet	-	94.4 · **98.7** · 99.7	93.6 · **98.2** · 99.3	89.4 · **96.3** · 99.8	84.5 · **95.1** · 100.0	81.7 · **92.5** · 100.0
	DDNet	-	95.5 · **98.6** · 99.0	94.6 · **98.7** · 99.4	89.7 · **97.1** · 99.8	80.0 · **93.6** · 100.0	54.4 · **74.5** · 100.0
	EvNet	-	88.9 · **94.8** · 98.1	91.5 · **98.4** · 99.8	89.2 · **97.0** · 100.0	83.6 · **94.7** · 100.0	72.3 · **88.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	92.8 · **98.5** · 99.9	92.8 · **98.7** · 99.9	89.0 · **96.3** · 99.8	80.8 · **93.4** · 100.0	71.6 · **86.9** · 100.0
	PriorNet	-	95.1 · **98.1** · 98.9	94.3 · **97.7** · 99.1	88.5 · **96.8** · 99.9	83.4 · **94.5** · 100.0	78.9 · **92.2** · 100.0
	DDNet	-	96.0 · **98.7** · 99.0	95.5 · **98.6** · 99.3	89.5 · **95.6** · 99.7	79.6 · **93.1** · 100.0	55.9 · **77.1** · 100.0
	EvNet	-	93.3 · **98.9** · 99.4	90.1 · **97.9** · 99.4	87.9 · **96.3** · 100.0	84.1 · **94.2** · 100.0	69.2 · **86.9** · 100.0

Table 35. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under FGSM label attacks. Smoothed DBU models on Sensorless. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	94.5 · **98.1** · 100.0	10.3 · **19.6** · 100.0	5.1 · **11.0** · 100.0	6.4 · **6.4** · 6.4	10.4 · **10.4** · 10.4	11.4 · **11.4** · 11.4
	PriorNet	97.1 · **99.5** · 100.0	13.6 · **27.3** · 100.0	6.3 · **12.4** · 100.0	2.6 · **6.8** · 100.0	3.1 · **7.3** · 100.0	3.2 · **6.7** · 100.0
	DDNet	95.9 · **99.4** · 99.8	8.6 · **14.9** · 100.0	1.6 · **3.8** · 100.0	2.6 · **4.5** · 100.0	3.4 · **6.9** · 100.0	3.3 · **6.4** · 100.0
	EvNet	94.0 · **98.5** · 99.7	26.0 · **43.2** · 100.0	15.8 · **30.8** · 100.0	11.7 · **20.2** · 100.0	8.1 · **15.0** · 100.0	7.6 · **12.7** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	13.1 · **24.3** · 100.0	5.7 · **11.9** · 100.0	9.4 · **9.4** · 9.4	11.2 · **11.2** · 11.2	11.8 · **11.8** · 11.8
	PriorNet	-	22.4 · **38.2** · 100.0	11.8 · **22.1** · 100.0	0.2 · **0.6** · 100.0	0.0 · **0.0** · 100.0	0.1 · **0.1** · 100.0
	DDNet	-	7.3 · **13.2** · 100.0	8.5 · **17.2** · 100.0	3.6 · **7.9** · 100.0	3.8 · **7.6** · 100.0	0.8 · **1.2** · 100.0
	EvNet	-	25.5 · **42.0** · 100.0	15.6 · **30.2** · 100.0	10.4 · **19.5** · 100.0	8.6 · **16.4** · 100.0	7.8 · **14.7** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	10.6 · **20.3** · 100.0	5.2 · **9.9** · 100.0	10.9 · **10.9** · 10.9	11.6 · **11.6** · 11.6	11.7 · **11.7** · 11.7
	PriorNet	-	25.7 · **45.0** · 100.0	12.0 · **20.5** · 100.0	1.1 · **3.7** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	-	7.9 · **16.4** · 100.0	1.2 · **3.8** · 100.0	3.4 · **6.3** · 100.0	3.9 · **7.9** · 100.0	3.3 · **8.0** · 100.0
	EvNet	-	27.9 · **49.2** · 100.0	18.4 · **32.9** · 100.0	16.4 · **29.3** · 100.0	5.9 · **10.8** · 100.0	8.5 · **16.1** · 100.0

Table 36. Distinguishing between correctly and wrongly labeled inputs based on differential entropy under FGSM label attacks. Smoothed DBU models on Segment. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	94.0 · **99.2** · 99.8	55.2 · **78.3** · 100.0	40.1 · **61.4** · 100.0	17.9 · **31.7** · 100.0	6.8 · **12.7** · 100.0	17.6 · **17.9** · 18.0
	PriorNet	97.0 · **99.8** · 99.9	69.2 · **89.7** · 100.0	29.7 · **45.5** · 100.0	1.7 · **4.1** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	96.2 · **99.5** · 99.6	70.6 · **86.3** · 99.8	22.3 · **38.8** · 100.0	6.3 · **13.3** · 100.0	1.1 · **3.0** · 100.0	0.0 · **0.0** · 100.0
	EvNet	95.8 · **99.1** · 99.8	78.4 · **92.5** · 100.0	40.7 · **62.1** · 100.0	9.8 · **17.6** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	66.0 · **83.5** · 100.0	28.8 · **44.9** · 100.0	12.3 · **24.3** · 100.0	9.3 · **17.3** · 100.0	24.8 · **24.8** · 24.8
	PriorNet	-	75.1 · **91.5** · 99.9	34.0 · **60.3** · 100.0	11.1 · **24.6** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	-	65.4 · **82.8** · 99.5	23.1 · **35.3** · 100.0	4.8 · **10.4** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	EvNet	-	83.4 · **95.3** · 100.0	42.1 · **63.3** · 100.0	15.0 · **33.6** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	67.8 · **86.5** · 100.0	34.0 · **52.5** · 100.0	16.2 · **32.8** · 100.0	14.4 · **25.2** · 92.2	7.3 · **7.3** · 7.3
	PriorNet	-	77.3 · **91.2** · 99.9	39.3 · **62.7** · 100.0	9.0 · **17.8** · 100.0	0.0 · **0.0** · 100.0	0.0 · **0.0** · 100.0
	DDNet	-	68.8 · **88.3** · 99.9	20.4 · **35.2** · 100.0	7.5 · **12.6** · 100.0	0.3 · **0.9** · 100.0	0.0 · **0.0** · 100.0
	EvNet	-	74.0 · **92.9** · 100.0	44.1 · **61.8** · 100.0	5.3 · **13.0** · 100.0	3.9 · **8.2** · 100.0	0.5 · **4.2** · 100.0

Table 37. Attack detection (PGD label attacks) based on differential entropy. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	33.1 · **50.4** · 89.9	31.0 · **50.2** · 96.9	30.7 · **50.2** · 100.0	30.7 · **50.0** · 100.0	30.7 · **50.2** · 100.0
	PriorNet	35.9 · **50.6** · 74.5	33.0 · **50.3** · 82.8	31.2 · **50.0** · 95.7	30.7 · **50.4** · 99.9	30.7 · **50.4** · 100.0
	DDNet	36.3 · **50.3** · 76.4	32.8 · **49.9** · 84.6	30.8 · **50.1** · 98.0	30.7 · **50.2** · 100.0	30.7 · **50.2** · 100.0
	EvNet	32.9 · **50.4** · 89.8	31.4 · **50.1** · 94.0	30.8 · **50.0** · 98.0	30.7 · **50.3** · 100.0	30.7 · **49.6** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	32.7 · **50.1** · 90.4	31.1 · **50.2** · 96.5	30.7 · **50.2** · 99.7	30.7 · **50.3** · 100.0	30.7 · **50.2** · 100.0
	PriorNet	35.2 · **51.8** · 78.6	32.8 · **51.1** · 84.4	30.8 · **50.2** · 98.7	30.7 · **50.5** · 100.0	30.8 · **50.1** · 98.2
	DDNet	35.5 · **50.6** · 79.2	33.4 · **50.3** · 84.1	30.8 · **50.1** · 99.2	30.7 · **50.0** · 100.0	30.7 · **50.5** · 100.0
	EvNet	40.3 · **50.4** · 66.8	31.4 · **50.3** · 95.8	30.7 · **50.3** · 100.0	30.7 · **50.1** · 100.0	30.7 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	33.3 · **50.6** · 88.7	32.5 · **50.1** · 87.9	30.7 · **49.9** · 99.8	30.7 · **50.1** · 100.0	30.7 · **50.0** · 100.0
	PriorNet	34.5 · **51.0** · 80.1	31.4 · **50.6** · 92.8	30.9 · **50.0** · 97.7	30.7 · **50.1** · 100.0	30.7 · **50.0** · 100.0
	DDNet	37.4 · **50.8** · 74.5	33.4 · **50.2** · 83.0	30.9 · **50.1** · 96.8	30.8 · **49.9** · 98.1	30.7 · **49.9** · 100.0
	EvNet	32.8 · **50.1** · 92.0	30.8 · **50.0** · 99.6	30.7 · **50.1** · 100.0	31.2 · **50.2** · 96.1	31.0 · **50.0** · 100.0

Table 38. Attack detection (PGD label attacks) based on differential entropy. Smoothed DBU models on MNIST. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	30.9 · **52.5** · 95.6	31.5 · **51.5** · 90.9	31.1 · **49.9** · 97.1	30.7 · **47.6** · 100.0	30.7 · **45.0** · 100.0
	PriorNet	38.2 · **57.8** · 80.9	36.0 · **57.2** · 84.3	31.6 · **63.4** · 98.4	30.8 · **61.0** · 99.3	30.7 · **66.8** · 100.0
	DDNet	44.6 · **51.9** · 60.7	39.3 · **52.7** · 72.2	31.6 · **50.9** · 95.2	30.7 · **47.3** · 100.0	30.7 · **45.9** · 100.0
	EvNet	36.5 · **51.8** · 76.1	31.5 · **51.1** · 93.2	30.7 · **51.1** · 99.9	30.7 · **48.7** · 100.0	30.7 · **43.8** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	33.6 · **52.8** · 82.3	31.4 · **51.2** · 91.6	30.9 · **49.4** · 99.1	30.7 · **49.3** · 100.0	30.7 · **56.0** · 100.0
	PriorNet	37.3 · **60.5** · 84.3	34.3 · **59.9** · 87.9	32.1 · **61.0** · 97.0	30.7 · **69.3** · 100.0	30.7 · **68.0** · 100.0
	DDNet	44.8 · **52.2** · 61.0	40.2 · **52.6** · 70.0	32.5 · **52.4** · 94.6	30.7 · **50.3** · 100.0	30.7 · **54.6** · 100.0
	EvNet	35.8 · **51.2** · 76.7	32.9 · **51.0** · 88.5	30.7 · **49.5** · 100.0	30.7 · **48.5** · 100.0	30.7 · **47.7** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	31.2 · **52.7** · 92.8	31.3 · **51.7** · 92.4	31.3 · **47.3** · 96.8	30.7 · **48.9** · 100.0	30.7 · **46.3** · 100.0
	PriorNet	38.3 · **58.2** · 81.5	36.9 · **55.5** · 79.9	31.3 · **63.5** · 98.9	30.7 · **68.6** · 100.0	30.7 · **74.6** · 100.0
	DDNet	44.9 · **52.2** · 60.7	39.6 · **53.3** · 72.1	31.8 · **51.7** · 95.4	30.7 · **46.1** · 100.0	30.7 · **46.0** · 100.0
	EvNet	38.8 · **51.9** · 70.9	34.5 · **52.3** · 82.9	30.8 · **49.9** · 99.6	30.7 · **47.7** · 100.0	30.8 · **49.4** · 100.0

Table 39. Attack detection (PGD label attacks) based on differential entropy. Smoothed DBU models on Sensorless. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	30.7 · **61.9** · 100.0	30.7 · **60.1** · 100.0	46.5 · **50.0** · 75.5	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	30.7 · **50.1** · 100.0	30.7 · **46.5** · 100.0	30.7 · **42.3** · 100.0	30.7 · **66.7** · 100.0	30.9 · **79.2** · 100.0
	DDNet	30.7 · **57.5** · 100.0	30.7 · **49.9** · 100.0	30.7 · **45.5** · 100.0	30.7 · **50.0** · 100.0	30.7 · **59.3** · 100.0
	EvNet	30.7 · **62.0** · 100.0	30.7 · **59.6** · 100.0	30.7 · **55.8** · 100.0	30.7 · **48.3** · 100.0	31.8 · **50.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	30.7 · **58.8** · 100.0	30.7 · **58.2** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	30.7 · **60.2** · 100.0	30.7 · **54.6** · 100.0	30.7 · **45.0** · 100.0	30.7 · **38.0** · 100.0	33.9 · **49.9** · 100.0
	DDNet	30.7 · **55.4** · 100.0	30.7 · **53.7** · 100.0	30.7 · **44.6** · 100.0	30.7 · **38.8** · 100.0	30.7 · **51.9** · 100.0
	EvNet	30.7 · **62.1** · 100.0	30.7 · **54.3** · 100.0	30.7 · **59.9** · 100.0	30.7 · **62.1** · 100.0	30.7 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	30.7 · **63.0** · 100.0	30.7 · **54.0** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	30.7 · **58.0** · 100.0	30.7 · **55.6** · 100.0	30.7 · **44.2** · 100.0	30.7 · **53.5** · 100.0	30.7 · **78.5** · 100.0
	DDNet	30.7 · **55.1** · 100.0	30.7 · **48.2** · 100.0	30.7 · **50.1** · 100.0	30.7 · **52.6** · 100.0	30.7 · **57.0** · 100.0
	EvNet	30.7 · **63.5** · 100.0	30.7 · **54.3** · 100.0	30.7 · **54.2** · 100.0	30.7 · **45.0** · 100.0	30.7 · **50.0** · 100.0

Table 40. Attack detection (PGD label attacks) based on differential entropy. Smoothed DBU models on Segment. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	30.8 · **73.5** · 100.0	30.8 · **59.9** · 100.0	30.8 · **60.3** · 100.0	30.8 · **50.2** · 100.0	49.5 · **50.0** · 50.0
	PriorNet	30.9 · **77.1** · 99.9	30.8 · **78.1** · 100.0	30.8 · **39.5** · 100.0	30.8 · **35.2** · 100.0	30.8 · **41.4** · 100.0
	DDNet	31.4 · **69.6** · 99.5	30.8 · **71.2** · 100.0	30.8 · **54.3** · 100.0	30.8 · **35.5** · 100.0	30.8 · **35.7** · 100.0
	EvNet	30.8 · **86.2** · 100.0	30.8 · **80.3** · 100.0	30.8 · **54.0** · 100.0	30.8 · **43.3** · 100.0	30.8 · **40.5** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	30.8 · **75.6** · 100.0	30.8 · **69.7** · 100.0	30.8 · **66.5** · 100.0	30.8 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	31.0 · **74.4** · 99.2	30.8 · **74.0** · 100.0	30.8 · **59.8** · 100.0	30.8 · **56.0** · 100.0	30.8 · **38.8** · 100.0
	DDNet	31.6 · **68.9** · 99.0	30.8 · **72.9** · 100.0	30.8 · **47.5** · 100.0	30.8 · **32.2** · 100.0	30.8 · **31.8** · 100.0
	EvNet	30.8 · **83.4** · 100.0	30.8 · **87.0** · 100.0	30.8 · **61.9** · 100.0	30.8 · **39.2** · 100.0	30.8 · **41.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	30.8 · **73.9** · 100.0	30.8 · **64.5** · 100.0	30.8 · **68.3** · 100.0	33.0 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	31.0 · **73.7** · 99.6	30.8 · **73.1** · 100.0	30.8 · **57.8** · 100.0	30.8 · **44.8** · 100.0	30.8 · **49.1** · 100.0
	DDNet	31.0 · **70.7** · 99.7	30.8 · **70.6** · 100.0	30.8 · **48.6** · 100.0	30.8 · **31.6** · 100.0	30.8 · **30.9** · 100.0
	EvNet	30.8 · **85.8** · 100.0	30.8 · **86.7** · 100.0	30.8 · **54.4** · 100.0	30.8 · **45.1** · 100.0	30.8 · **34.8** · 100.0

Table 41. Attack detection (FGSM label attacks) based on differential entropy. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	33.1 · **50.3** · 89.9	31.0 · **50.2** · 96.9	30.7 · **50.1** · 100.0	30.7 · **49.5** · 100.0	30.7 · **50.2** · 100.0
	PriorNet	36.0 · **50.8** · 74.6	33.0 · **50.4** · 82.8	31.2 · **50.2** · 95.6	30.7 · **50.7** · 99.9	30.7 · **51.4** · 100.0
	DDNet	36.4 · **50.4** · 76.4	32.8 · **49.9** · 84.6	30.8 · **50.1** · 97.9	30.7 · **50.2** · 100.0	30.7 · **49.9** · 100.0
	EvNet	32.9 · **50.3** · 89.7	31.4 · **50.2** · 94.0	30.8 · **50.1** · 98.0	30.7 · **49.7** · 100.0	30.7 · **49.7** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	32.7 · **50.1** · 90.3	31.1 · **50.3** · 96.4	30.7 · **50.1** · 99.7	30.7 · **49.8** · 100.0	30.7 · **50.5** · 100.0
	PriorNet	35.4 · **52.3** · 78.9	32.9 · **51.3** · 84.5	30.7 · **50.3** · 98.7	30.7 · **50.7** · 100.0	30.8 · **50.2** · 98.2
	DDNet	35.5 · **50.6** · 79.3	33.4 · **50.3** · 84.2	30.8 · **50.1** · 99.2	30.7 · **49.9** · 100.0	30.7 · **50.1** · 100.0
	EvNet	40.3 · **50.4** · 66.8	31.4 · **50.3** · 95.9	30.7 · **50.2** · 100.0	30.7 · **50.1** · 100.0	30.7 · **49.6** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	33.3 · **50.7** · 88.7	32.5 · **50.1** · 87.8	30.7 · **50.1** · 99.8	30.7 · **50.5** · 100.0	30.7 · **50.2** · 100.0
	PriorNet	34.6 · **51.2** · 80.3	31.4 · **50.7** · 92.8	30.9 · **50.2** · 97.7	30.7 · **50.0** · 100.0	30.7 · **50.1** · 100.0
	DDNet	37.4 · **51.0** · 74.7	33.4 · **50.2** · 83.0	30.9 · **50.1** · 96.9	30.8 · **50.1** · 98.1	30.7 · **49.9** · 100.0
	EvNet	32.8 · **50.1** · 92.0	30.8 · **50.2** · 99.6	30.7 · **50.4** · 100.0	31.2 · **50.2** · 96.0	31.0 · **50.0** · 100.0

Table 42. Attack detection (FGSM label attacks) based on differential entropy. Smoothed DBU models on MNIST. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	30.9 · **52.3** · 95.6	31.5 · **51.2** · 90.8	31.1 · **49.8** · 97.0	30.7 · **48.3** · 100.0	30.7 · **46.5** · 100.0
	PriorNet	38.1 · **57.7** · 80.8	35.8 · **56.6** · 84.0	31.5 · **61.7** · 98.3	30.8 · **58.9** · 99.2	30.7 · **62.3** · 100.0
	DDNet	44.7 · **52.0** · 60.9	39.4 · **52.9** · 72.5	31.6 · **50.8** · 95.2	30.7 · **47.5** · 100.0	30.7 · **46.8** · 100.0
	EvNet	36.5 · **51.7** · 76.0	31.5 · **51.1** · 93.2	30.7 · **50.9** · 99.9	30.7 · **48.9** · 100.0	30.7 · **46.2** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	33.5 · **52.6** · 82.2	31.4 · **51.0** · 91.5	30.9 · **49.8** · 99.0	30.7 · **50.1** · 100.0	30.7 · **54.4** · 100.0
	PriorNet	37.3 · **60.6** · 84.3	34.2 · **59.5** · 87.8	32.1 · **60.0** · 96.9	30.7 · **66.3** · 100.0	30.7 · **63.3** · 100.0
	DDNet	44.9 · **52.3** · 61.0	40.3 · **52.8** · 70.2	32.5 · **52.4** · 94.6	30.7 · **50.0** · 100.0	30.7 · **57.1** · 100.0
	EvNet	35.8 · **51.5** · 76.7	32.9 · **50.9** · 88.5	30.7 · **50.0** · 100.0	30.7 · **48.9** · 100.0	30.7 · **48.6** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	31.2 · **52.6** · 92.9	31.3 · **51.5** · 92.3	31.3 · **48.0** · 96.9	30.7 · **49.4** · 100.0	30.7 · **48.1** · 100.0
	PriorNet	38.3 · **58.3** · 81.4	36.8 · **55.2** · 79.8	31.3 · **62.5** · 98.9	30.7 · **64.5** · 100.0	30.7 · **68.7** · 100.0
	DDNet	45.0 · **52.3** · 60.9	39.7 · **53.5** · 72.4	31.8 · **51.7** · 95.4	30.7 · **46.6** · 100.0	30.7 · **44.4** · 100.0
	EvNet	38.8 · **51.8** · 70.8	34.5 · **52.0** · 82.7	30.8 · **50.0** · 99.6	30.7 · **49.3** · 100.0	30.8 · **50.3** · 100.0

Table 43. Attack detection (FGSM label attacks) based on differential entropy. Smoothed DBU models on Sensorless. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	30.7 · **82.0** · 100.0	30.7 · **88.6** · 100.0	50.0 · **50.0** · 50.1	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	30.7 · **51.7** · 100.0	30.7 · **48.2** · 100.0	30.7 · **48.6** · 100.0	30.7 · **68.6** · 100.0	31.4 · **63.7** · 100.0
	DDNet	30.7 · **67.1** · 100.0	30.7 · **58.2** · 100.0	30.7 · **51.7** · 100.0	30.7 · **69.8** · 100.0	30.7 · **73.7** · 100.0
	EvNet	30.7 · **77.9** · 100.0	30.7 · **85.3** · 100.0	30.7 · **90.5** · 100.0	30.8 · **84.3** · 100.0	34.0 · **50.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	30.7 · **76.7** · 100.0	30.7 · **78.7** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	30.7 · **63.9** · 100.0	30.7 · **58.4** · 100.0	30.7 · **45.2** · 100.0	30.7 · **43.3** · 100.0	32.9 · **35.5** · 100.0
	DDNet	30.7 · **58.5** · 100.0	30.7 · **75.8** · 100.0	30.7 · **72.6** · 100.0	30.7 · **35.6** · 100.0	30.7 · **71.5** · 100.0
	EvNet	30.7 · **80.4** · 100.0	30.7 · **71.5** · 100.0	30.7 · **75.3** · 100.0	30.7 · **78.5** · 100.0	30.7 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	30.7 · **77.4** · 100.0	30.7 · **68.0** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	30.7 · **63.8** · 100.0	30.7 · **64.1** · 100.0	30.7 · **46.9** · 100.0	30.7 · **48.9** · 100.0	30.7 · **78.0** · 100.0
	DDNet	30.7 · **56.5** · 100.0	30.7 · **54.6** · 100.0	30.7 · **59.4** · 100.0	30.7 · **71.8** · 100.0	30.7 · **76.0** · 100.0
	EvNet	30.7 · **71.5** · 100.0	30.7 · **75.7** · 100.0	30.7 · **90.5** · 100.0	30.7 · **54.7** · 100.0	30.9 · **50.2** · 100.0

Table 44. Attack detection (FGSM label attacks) based on differential entropy. Smoothed DBU models on Segment. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model)..

	Att. Rad.	0.1	0.2	0.5	1.0	2.0
Smoothed models	**PostNet**	30.8 · **76.9** · 100.0	30.8 · **62.5** · 100.0	30.8 · **59.2** · 100.0	30.8 · **48.7** · 100.0	49.7 · **50.0** · 50.0
	PriorNet	30.9 · **81.3** · 99.9	30.8 · **85.0** · 100.0	30.8 · **48.7** · 100.0	30.8 · **37.1** · 100.0	30.8 · **43.7** · 100.0
	DDNet	31.7 · **73.8** · 99.7	30.8 · **80.5** · 100.0	30.8 · **80.4** · 100.0	30.8 · **72.7** · 100.0	30.8 · **70.6** · 100.0
	EvNet	30.8 · **89.1** · 100.0	30.8 · **89.5** · 100.0	30.8 · **75.3** · 100.0	30.8 · **73.1** · 100.0	30.8 · **83.1** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	30.8 · **81.0** · 100.0	30.8 · **75.6** · 100.0	30.8 · **56.3** · 100.0	30.8 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	31.1 · **77.9** · 99.4	30.8 · **76.1** · 100.0	30.8 · **62.4** · 100.0	30.8 · **65.5** · 100.0	30.8 · **53.3** · 100.0
	DDNet	31.9 · **72.5** · 99.3	30.8 · **82.0** · 100.0	30.8 · **65.7** · 100.0	30.8 · **53.0** · 100.0	30.8 · **61.6** · 100.0
	EvNet	30.8 · **86.4** · 100.0	30.8 · **94.1** · 100.0	30.8 · **78.6** · 100.0	30.8 · **77.7** · 100.0	30.8 · **85.5** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	30.8 · **76.8** · 100.0	30.8 · **64.6** · 100.0	30.8 · **82.9** · 100.0	32.2 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	31.1 · **77.6** · 99.7	30.8 · **76.7** · 100.0	30.8 · **69.0** · 100.0	30.8 · **53.1** · 100.0	30.8 · **61.4** · 100.0
	DDNet	31.1 · **74.3** · 99.8	30.8 · **77.1** · 100.0	30.8 · **76.0** · 100.0	30.8 · **57.0** · 100.0	30.8 · **43.5** · 100.0
	EvNet	30.8 · **88.8** · 100.0	30.8 · **92.6** · 100.0	30.8 · **70.2** · 100.0	30.8 · **62.0** · 100.0	30.8 · **96.2** · 100.0

Table 45. OOD detection based on differential entropy under PGD uncertainty attacks against differential entorpy on ID data and OOD data. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
				ID-Attack			
Smoothed models	**PostNet**	72.1 · **82.7** · 88.0	35.0 · **56.6** · 97.4	31.9 · **65.6** · 99.8	30.7 · **50.6** · 100.0	30.7 · **46.9** · 100.0	30.7 · **51.6** · 100.0
	PriorNet	50.2 · **53.1** · 55.9	33.5 · **43.3** · 65.3	31.3 · **39.7** · 69.1	31.3 · **48.3** · 98.2	30.7 · **44.4** · 99.9	30.7 · **45.4** · 100.0
	DDNet	72.0 · **75.8** · 79.8	35.6 · **46.2** · 69.8	32.9 · **50.3** · 87.1	31.1 · **58.7** · 98.6	30.7 · **59.3** · 100.0	30.7 · **44.5** · 100.0
	EvNet	79.5 · **87.1** · 92.8	34.1 · **58.6** · 95.1	32.5 · **61.2** · 96.9	31.7 · **60.6** · 98.7	30.7 · **62.4** · 100.0	30.7 · **57.3** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	35.0 · **58.5** · 97.7	31.2 · **46.6** · 97.4	30.8 · **57.7** · 99.7	30.7 · **49.8** · 100.0	30.7 · **50.9** · 100.0
	PriorNet	-	31.5 · **36.7** · 57.2	33.1 · **51.8** · 84.8	30.7 · **57.7** · 98.7	30.7 · **40.0** · 99.9	30.9 · **53.6** · 96.7
	DDNet	-	36.2 · **50.0** · 78.6	32.1 · **41.3** · 70.2	30.8 · **56.4** · 100.0	30.7 · **49.4** · 100.0	30.7 · **54.8** · 100.0
	EvNet	-	46.8 · **61.0** · 79.7	32.3 · **58.9** · 99.1	30.7 · **45.0** · 100.0	30.7 · **63.3** · 100.0	30.8 · **38.1** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	35.2 · **55.9** · 96.0	34.5 · **59.2** · 94.9	30.7 · **47.0** · 100.0	30.7 · **58.2** · 100.0	30.7 · **42.9** · 100.0
	PriorNet	-	31.8 · **38.9** · 64.1	31.0 · **41.8** · 87.9	30.7 · **42.9** · 99.2	30.7 · **48.6** · 100.0	30.7 · **46.6** · 100.0
	DDNet	-	39.7 · **52.1** · 75.7	36.4 · **56.8** · 83.8	31.0 · **51.5** · 97.4	31.0 · **56.8** · 97.8	30.7 · **49.1** · 100.0
	EvNet	-	34.8 · **64.9** · 99.6	30.8 · **48.9** · 99.8	30.7 · **66.8** · 100.0	30.9 · **41.5** · 93.8	31.1 · **55.1** · 100.0
				OOD-Attack			
Smoothed models	**PostNet**	72.0 · **82.7** · 88.0	35.1 · **56.8** · 97.3	32.0 · **65.8** · 99.8	30.7 · **50.7** · 100.0	30.7 · **46.5** · 100.0	30.7 · **51.7** · 100.0
	PriorNet	50.3 · **53.1** · 55.9	33.6 · **43.7** · 65.9	31.3 · **39.8** · 69.4	31.3 · **48.3** · 98.2	30.7 · **44.5** · 99.9	30.7 · **46.4** · 100.0
	DDNet	72.0 · **75.8** · 79.8	35.6 · **46.2** · 70.0	32.9 · **50.1** · 86.7	31.1 · **58.8** · 98.6	30.7 · **59.3** · 100.0	30.7 · **44.6** · 100.0
	EvNet	79.5 · **87.1** · 92.8	34.1 · **58.8** · 95.2	32.6 · **61.2** · 96.9	31.7 · **60.5** · 98.7	30.7 · **62.4** · 100.0	30.7 · **57.6** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	35.0 · **58.5** · 97.8	31.2 · **46.6** · 97.2	30.8 · **57.7** · 99.7	30.7 · **50.2** · 100.0	30.7 · **51.5** · 100.0
	PriorNet	-	31.6 · **37.3** · 59.3	33.2 · **52.7** · 85.8	30.7 · **57.8** · 98.7	30.7 · **40.1** · 99.9	30.9 · **53.8** · 96.8
	DDNet	-	36.4 · **50.2** · 78.9	32.1 · **41.5** · 70.4	30.9 · **56.2** · 100.0	30.7 · **49.3** · 100.0	30.7 · **55.1** · 100.0
	EvNet	-	47.2 · **61.1** · 80.0	32.4 · **59.1** · 99.1	30.7 · **45.0** · 100.0	30.7 · **63.2** · 100.0	30.8 · **38.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	35.3 · **56.4** · 96.1	34.5 · **59.0** · 94.9	30.7 · **46.8** · 100.0	30.7 · **57.8** · 100.0	30.7 · **43.2** · 100.0
	PriorNet	-	31.9 · **39.4** · 65.5	31.0 · **42.0** · 88.6	30.7 · **42.9** · 99.2	30.7 · **48.4** · 100.0	30.7 · **47.1** · 100.0
	DDNet	-	40.2 · **52.9** · 76.5	36.4 · **56.9** · 83.9	31.1 · **51.5** · 97.3	31.0 · **57.0** · 97.8	30.7 · **49.1** · 100.0
	EvNet	-	34.9 · **64.8** · 99.6	30.8 · **48.8** · 99.8	30.7 · **66.1** · 100.0	30.9 · **41.6** · 93.6	31.1 · **54.7** · 100.0

Table 46. OOD detection based on differential entropy under PGD uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on MNIST. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
				ID-Attack			
	PostNet	59.9 · **91.1** · 98.6	61.2 · **97.7** · 99.6	64.8 · **94.7** · 99.7	31.6 · **64.9** · 99.7	30.7 · **63.2** · 100.0	30.7 · **70.5** · 100.0
Smoothed	**PriorNet**	99.8 · **99.8** · 99.8	99.4 · **99.8** · 99.9	98.3 · **99.6** · 99.9	48.5 · **91.9** · 99.9	31.1 · **74.6** · 99.8	30.7 · **67.3** · 100.0
models	**DDNet**	98.5 · **98.6** · 98.7	95.0 · **97.6** · 98.9	74.7 · **92.0** · 98.2	31.4 · **52.0** · 98.5	30.7 · **52.0** · 100.0	30.7 · **41.1** · 100.0
	EvNet	85.7 · **87.5** · 89.2	68.9 · **90.4** · 97.7	42.5 · **90.2** · 99.6	30.7 · **69.8** · 100.0	30.7 · **50.3** · 100.0	30.7 · **45.6** · 100.0
Smoothed	**PostNet**	-	84.3 · **96.2** · 99.3	50.4 · **89.2** · 99.5	30.9 · **46.2** · 99.4	30.7 · **46.9** · 100.0	30.7 · **62.2** · 100.0
+ adv. w.	**PriorNet**	-	99.7 · **99.9** · 100.0	98.7 · **99.8** · 100.0	83.3 · **99.1** · 100.0	30.7 · **82.6** · 100.0	30.7 · **64.8** · 100.0
label	**DDNet**	-	93.6 · **96.9** · 98.5	71.2 · **89.1** · 96.9	32.3 · **50.3** · 99.0	30.7 · **50.7** · 100.0	30.7 · **55.7** · 100.0
attacks	**EvNet**	-	58.2 · **84.4** · 94.3	40.9 · **87.4** · 99.2	30.7 · **59.4** · 100.0	30.7 · **40.3** · 100.0	30.7 · **53.2** · 100.0
Smoothed	**PostNet**	-	58.9 · **96.1** · 99.3	59.7 · **96.1** · 99.9	31.2 · **48.2** · 95.7	30.7 · **42.0** · 100.0	30.7 · **56.9** · 100.0
+ adv. w.	**PriorNet**	-	99.9 · **100.0** · 100.0	96.5 · **99.2** · 99.9	49.2 · **96.9** · 100.0	31.3 · **88.1** · 100.0	30.7 · **77.8** · 100.0
uncert.	**DDNet**	-	95.0 · **97.5** · 98.8	80.6 · **94.1** · 98.7	31.7 · **55.6** · 98.6	30.7 · **52.0** · 100.0	30.7 · **47.6** · 100.0
attacks	**EvNet**	-	66.5 · **91.3** · 98.1	48.1 · **84.1** · 97.6	30.8 · **49.7** · 99.9	30.7 · **37.9** · 100.0	30.8 · **63.5** · 100.0
				OOD-Attack			
	PostNet	59.0 · **91.2** · 97.7	57.8 · **97.2** · 99.6	61.4 · **93.8** · 99.6	31.5 · **58.9** · 99.5	30.7 · **51.5** · 100.0	30.7 · **53.5** · 100.0
Smoothed	**PriorNet**	99.7 · **99.8** · 99.8	99.4 · **99.8** · 99.9	98.4 · **99.7** · 100.0	60.7 · **96.8** · 100.0	33.0 · **88.9** · 100.0	30.7 · **87.7** · 100.0
models	**DDNet**	98.4 · **98.5** · 98.7	94.2 · **97.2** · 98.7	72.1 · **90.5** · 97.8	31.6 · **52.3** · 98.1	30.7 · **51.7** · 100.0	30.7 · **37.7** · 100.0
	EvNet	83.9 · **85.7** · 88.0	63.5 · **88.6** · 97.9	40.1 · **87.7** · 99.6	30.8 · **68.9** · 100.0	30.7 · **43.3** · 100.0	30.7 · **36.8** · 100.0
Smoothed	**PostNet**	-	84.7 · **96.1** · 99.4	49.7 · **89.1** · 99.5	30.9 · **45.6** · 99.3	30.7 · **45.8** · 100.0	30.7 · **69.1** · 100.0
+ adv. w.	**PriorNet**	-	99.7 · **99.9** · 100.0	98.7 · **99.8** · 100.0	86.8 · **99.5** · 100.0	30.9 · **93.2** · 100.0	30.7 · **81.4** · 100.0
label	**DDNet**	-	93.9 · **97.0** · 98.6	72.0 · **89.4** · 97.0	33.0 · **52.4** · 98.8	30.7 · **51.5** · 100.0	30.7 · **60.1** · 100.0
attacks	**EvNet**	-	59.5 · **85.3** · 94.6	40.7 · **86.9** · 99.2	30.7 · **57.4** · 100.0	30.7 · **39.2** · 100.0	30.7 · **49.0** · 100.0
Smoothed	**PostNet**	-	55.7 · **96.1** · 99.3	58.4 · **95.7** · 99.8	31.1 · **44.2** · 93.1	30.7 · **41.2** · 100.0	30.7 · **48.8** · 100.0
+ adv. w.	**PriorNet**	-	99.9 · **100.0** · 100.0	97.0 · **99.3** · 99.9	61.0 · **98.4** · 100.0	33.2 · **94.4** · 100.0	30.7 · **90.2** · 100.0
uncert.	**DDNet**	-	95.3 · **97.6** · 98.9	82.2 · **94.5** · 98.7	32.1 · **56.6** · 98.5	30.7 · **48.6** · 100.0	30.7 · **42.9** · 100.0
attacks	**EvNet**	-	65.2 · **90.4** · 98.0	46.8 · **83.4** · 97.3	30.8 · **48.8** · 99.9	30.7 · **36.3** · 100.0	30.8 · **60.1** · 100.0

Table 47. OOD detection based on differential entropy under PGD uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on Sensorless. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
				ID-Attack			
Smoothed models	**PostNet**	49.3 · **90.4** · 99.8	30.7 · **49.2** · 100.0	30.7 · **36.0** · 100.0	49.2 · **50.0** · 74.9	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	31.2 · **39.0** · 66.9	30.7 · **35.5** · 100.0	30.7 · **38.9** · 100.0	30.7 · **46.2** · 100.0	30.7 · **62.7** · 100.0	30.7 · **51.3** · 100.0
	DDNet	31.0 · **31.5** · 32.7	30.7 · **30.8** · 100.0	30.7 · **31.8** · 100.0	30.7 · **53.6** · 100.0	30.7 · **43.9** · 100.0	30.7 · **40.5** · 100.0
	EvNet	33.6 · **55.2** · 91.3	30.7 · **44.2** · 100.0	30.7 · **43.8** · 100.0	30.7 · **39.3** · 100.0	30.8 · **51.6** · 100.0	32.4 · **50.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	30.7 · **62.4** · 100.0	30.7 · **39.2** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **30.9** · 100.0	30.7 · **32.4** · 100.0	30.7 · **31.0** · 100.0	30.8 · **30.7** · 100.0	38.2 · **48.9** · 100.0
	DDNet	-	30.7 · **32.2** · 100.0	30.7 · **30.9** · 100.0	30.7 · **37.1** · 100.0	30.7 · **42.1** · 100.0	30.7 · **37.7** · 100.0
	EvNet	-	30.7 · **48.9** · 100.0	30.7 · **34.0** · 100.0	30.7 · **35.6** · 100.0	30.7 · **33.6** · 100.0	30.7 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.7 · **46.0** · 100.0	30.7 · **46.6** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **35.8** · 100.0	30.7 · **32.1** · 100.0	30.7 · **81.6** · 100.0	30.8 · **41.7** · 100.0	30.7 · **61.9** · 100.0
	DDNet	-	30.7 · **32.8** · 100.0	30.7 · **31.0** · 100.0	30.7 · **31.8** · 100.0	30.7 · **43.7** · 100.0	30.7 · **34.7** · 100.0
	EvNet	-	30.7 · **31.0** · 100.0	30.7 · **49.6** · 100.0	30.7 · **47.7** · 100.0	30.7 · **42.6** · 100.0	30.7 · **50.0** · 100.0
				OOD-Attack			
Smoothed models	**PostNet**	49.3 · **90.4** · 99.8	30.8 · **76.4** · 100.0	30.7 · **61.3** · 100.0	47.7 · **50.0** · 75.1	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	31.2 · **39.0** · 66.9	30.7 · **33.9** · 100.0	30.7 · **34.3** · 100.0	30.7 · **37.0** · 100.0	30.7 · **74.0** · 100.0	30.9 · **78.1** · 100.0
	DDNet	31.0 · **31.5** · 32.7	30.7 · **30.7** · 100.0	30.7 · **31.8** · 100.0	30.7 · **47.7** · 100.0	30.7 · **43.8** · 100.0	30.7 · **52.5** · 100.0
	EvNet	33.6 · **55.2** · 91.2	30.7 · **54.7** · 100.0	30.7 · **54.0** · 100.0	30.7 · **51.0** · 100.0	30.7 · **45.2** · 100.0	31.7 · **50.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	30.7 · **82.2** · 100.0	30.7 · **61.4** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **31.2** · 100.0	30.7 · **31.4** · 99.9	30.7 · **30.8** · 100.0	30.8 · **30.7** · 100.0	33.8 · **34.0** · 100.0
	DDNet	-	30.7 · **32.2** · 100.0	30.7 · **30.8** · 100.0	30.7 · **33.6** · 100.0	30.7 · **46.9** · 100.0	30.7 · **40.3** · 100.0
	EvNet	-	30.8 · **75.3** · 100.0	30.7 · **31.6** · 100.0	30.7 · **42.1** · 100.0	30.7 · **38.7** · 100.0	30.7 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.7 · **73.7** · 100.0	30.7 · **61.6** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **35.9** · 100.0	30.7 · **30.7** · 100.0	30.7 · **39.4** · 100.0	30.7 · **36.6** · 100.0	30.7 · **97.6** · 100.0
	DDNet	-	30.7 · **32.1** · 100.0	30.7 · **30.8** · 100.0	30.7 · **32.2** · 100.0	30.7 · **50.7** · 100.0	30.7 · **39.8** · 100.0
	EvNet	-	30.7 · **31.3** · 100.0	30.8 · **39.7** · 100.0	30.7 · **52.2** · 100.0	30.7 · **42.3** · 100.0	30.7 · **50.0** · 100.0

Table 48. OOD detection based on differential entropy under PGD uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on Segment. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
				ID-Attack			
Smoothed models	**PostNet**	99.6 · **99.9** · 99.9	33.0 · **83.0** · 100.0	30.8 · **43.8** · 100.0	30.8 · **31.7** · 100.0	30.8 · **40.8** · 100.0	41.4 · **50.0** · 50.2
	PriorNet	30.8 · **31.0** · 31.4	30.8 · **30.8** · 42.6	30.8 · **30.8** · 95.5	30.8 · **33.1** · 100.0	30.8 · **76.4** · 100.0	30.8 · **78.7** · 100.0
	DDNet	30.8 · **30.8** · 30.8	30.8 · **30.8** · 32.1	30.8 · **30.8** · 69.4	30.8 · **30.8** · 100.0	30.8 · **31.0** · 100.0	30.8 · **33.4** · 100.0
	EvNet	94.9 · **97.2** · 98.3	31.1 · **75.8** · 99.9	30.8 · **74.2** · 100.0	30.8 · **62.9** · 100.0	30.8 · **58.1** · 100.0	30.8 · **43.4** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	31.0 · **70.9** · 100.0	30.8 · **47.1** · 100.0	30.8 · **85.0** · 100.0	30.8 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **30.8** · 46.0	30.8 · **30.8** · 32.7	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.9 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 30.8	30.8 · **30.8** · 79.5	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **57.3** · 100.0
	EvNet	-	36.3 · **94.3** · 100.0	30.8 · **32.2** · 100.0	30.8 · **50.2** · 100.0	30.8 · **93.9** · 100.0	30.8 · **56.3** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.8 · **49.5** · 100.0	30.8 · **34.5** · 100.0	30.8 · **96.1** · 100.0	41.2 · **50.0** · 82.7	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **31.2** · 62.6	30.8 · **30.8** · 32.9	30.8 · **30.8** · 88.9	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 31.2	30.8 · **30.8** · 68.9	30.8 · **30.8** · 100.0	30.8 · **30.9** · 100.0	30.8 · **38.6** · 100.0
	EvNet	-	30.9 · **83.5** · 100.0	30.8 · **84.0** · 100.0	30.8 · **98.6** · 100.0	30.8 · **92.8** · 100.0	30.8 · **45.6** · 100.0
				OOD-Attack			
Smoothed models	**PostNet**	99.6 · **99.9** · 99.9	31.3 · **95.2** · 100.0	30.8 · **48.7** · 100.0	30.8 · **34.0** · 100.0	30.8 · **41.0** · 100.0	41.8 · **50.0** · 50.2
	PriorNet	30.8 · **31.0** · 31.4	30.8 · **30.8** · 44.7	30.8 · **30.8** · 86.3	30.8 · **30.9** · 100.0	30.8 · **35.7** · 100.0	30.8 · **57.4** · 100.0
	DDNet	30.8 · **30.8** · 30.8	30.8 · **30.8** · 31.9	30.8 · **30.8** · 58.3	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	EvNet	94.9 · **97.2** · 98.3	31.4 · **92.5** · 100.0	30.8 · **94.2** · 100.0	30.8 · **80.4** · 100.0	30.8 · **70.2** · 100.0	30.8 · **48.2** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	30.8 · **88.7** · 100.0	30.8 · **70.9** · 100.0	30.8 · **97.2** · 100.0	30.8 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **30.9** · 47.2	30.8 · **30.8** · 32.5	30.8 · **30.8** · 96.2	30.8 · **30.8** · 100.0	30.9 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 30.8	30.8 · **30.8** · 73.5	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **34.3** · 100.0
	EvNet	-	35.9 · **95.9** · 100.0	30.8 · **36.6** · 100.0	30.8 · **45.8** · 100.0	30.8 · **75.2** · 100.0	30.8 · **93.8** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.8 · **64.6** · 100.0	30.8 · **31.9** · 100.0	30.8 · **99.1** · 100.0	37.2 · **50.0** · 100.0	49.8 · **50.0** · 50.0
	PriorNet	-	30.8 · **31.3** · 60.6	30.8 · **30.8** · 34.8	30.8 · **30.8** · 73.8	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 31.7	30.8 · **30.8** · 64.6	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	EvNet	-	31.1 · **90.7** · 100.0	30.8 · **96.6** · 100.0	30.8 · **98.9** · 100.0	30.8 · **97.5** · 100.0	30.8 · **34.2** · 100.0

Table 49. OOD detection based on differential entropy under FGSM uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on CIFAR10. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
				ID-Attack			
Smoothed models	**PostNet**	72.2 · **82.7** · 88.0	35.0 · **56.5** · 97.5	31.9 · **65.5** · 99.8	30.7 · **50.6** · 100.0	30.7 · **46.9** · 100.0	30.7 · **51.4** · 100.0
	PriorNet	50.3 · **53.1** · 55.9	33.5 · **43.2** · 65.0	31.3 · **39.7** · 69.1	31.3 · **48.3** · 98.2	30.7 · **44.2** · 99.9	30.7 · **44.9** · 100.0
	DDNet	72.0 · **75.8** · 79.8	35.5 · **46.2** · 69.7	32.9 · **50.3** · 87.0	31.1 · **58.6** · 98.6	30.7 · **59.4** · 100.0	30.7 · **44.5** · 100.0
	EvNet	79.5 · **87.1** · 92.8	34.1 · **58.6** · 95.2	32.5 · **61.1** · 96.9	31.7 · **60.6** · 98.8	30.7 · **62.6** · 100.0	30.7 · **57.3** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	35.0 · **58.5** · 97.7	31.2 · **46.6** · 97.4	30.8 · **57.7** · 99.7	30.7 · **50.1** · 100.0	30.7 · **50.6** · 100.0
	PriorNet	-	31.5 · **36.6** · 56.7	33.1 · **51.7** · 84.4	30.7 · **57.5** · 98.7	30.7 · **40.1** · 99.9	30.9 · **53.5** · 96.7
	DDNet	-	36.2 · **50.0** · 78.5	32.1 · **41.3** · 70.1	30.9 · **56.3** · 100.0	30.7 · **49.5** · 100.0	30.7 · **54.9** · 100.0
	EvNet	-	46.8 · **60.9** · 79.6	32.3 · **58.9** · 99.1	30.7 · **45.1** · 100.0	30.7 · **63.1** · 100.0	30.8 · **38.1** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	35.2 · **56.0** · 95.9	34.5 · **59.0** · 94.8	30.7 · **47.0** · 100.0	30.7 · **57.2** · 100.0	30.7 · **42.7** · 100.0
	PriorNet	-	31.8 · **38.8** · 64.0	31.0 · **41.7** · 87.4	30.7 · **42.9** · 99.3	30.7 · **48.5** · 100.0	30.7 · **46.8** · 100.0
	DDNet	-	39.6 · **52.0** · 75.6	36.4 · **56.8** · 83.8	31.0 · **51.4** · 97.3	31.0 · **56.9** · 97.7	30.7 · **49.2** · 100.0
	EvNet	-	34.8 · **64.9** · 99.7	30.8 · **48.9** · 99.8	30.7 · **66.4** · 100.0	30.9 · **41.6** · 93.6	31.1 · **55.7** · 100.0
				OOD-Attack			
Smoothed models	**PostNet**	72.1 · **82.7** · 88.0	35.1 · **56.8** · 97.3	31.9 · **65.8** · 99.8	30.7 · **50.8** · 100.0	30.7 · **46.5** · 100.0	30.7 · **51.5** · 100.0
	PriorNet	50.3 · **53.1** · 55.9	33.6 · **43.7** · 65.9	31.3 · **39.8** · 69.4	31.3 · **48.3** · 98.2	30.7 · **44.4** · 99.9	30.7 · **45.9** · 100.0
	DDNet	72.0 · **75.8** · 79.8	35.6 · **46.1** · 70.0	32.9 · **50.1** · 86.7	31.1 · **58.7** · 98.6	30.7 · **59.3** · 100.0	30.7 · **44.6** · 100.0
	EvNet	79.5 · **87.1** · 92.8	34.1 · **58.8** · 95.2	32.6 · **61.3** · 96.9	31.7 · **60.5** · 98.8	30.7 · **62.2** · 100.0	30.7 · **57.7** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	35.0 · **58.4** · 97.9	31.2 · **46.6** · 97.3	30.8 · **57.7** · 99.7	30.7 · **50.1** · 100.0	30.7 · **51.4** · 100.0
	PriorNet	-	31.6 · **37.3** · 59.2	33.2 · **52.6** · 85.8	30.7 · **57.8** · 98.7	30.7 · **39.8** · 99.9	30.9 · **53.7** · 96.8
	DDNet	-	36.4 · **50.2** · 78.8	32.1 · **41.5** · 70.5	30.8 · **56.2** · 100.0	30.7 · **49.2** · 100.0	30.7 · **55.0** · 100.0
	EvNet	-	47.2 · **61.0** · 79.9	32.4 · **59.1** · 99.1	30.7 · **45.1** · 100.0	30.7 · **63.1** · 100.0	30.8 · **38.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	35.3 · **56.3** · 96.1	34.5 · **59.1** · 94.9	30.7 · **46.9** · 100.0	30.7 · **57.8** · 100.0	30.7 · **43.1** · 100.0
	PriorNet	-	31.9 · **39.4** · 65.4	31.0 · **42.0** · 88.7	30.7 · **42.9** · 99.2	30.7 · **48.3** · 100.0	30.7 · **47.2** · 100.0
	DDNet	-	40.1 · **52.8** · 76.5	36.5 · **56.9** · 83.9	31.1 · **51.5** · 97.3	31.0 · **57.0** · 97.8	30.7 · **48.7** · 100.0
	EvNet	-	34.9 · **65.0** · 99.6	30.8 · **48.8** · 99.8	30.7 · **66.6** · 100.0	30.9 · **41.1** · 93.4	31.1 · **55.3** · 100.0

Table 50. OOD detection based on differential entropy under FGSM uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on MNIST. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
					ID-Attack		
Smoothed models	**PostNet**	59.9 · **91.3** · 98.6	61.1 · **97.7** · 99.7	65.1 · **94.8** · 99.7	31.6 · **64.8** · 99.7	30.7 · **62.4** · 100.0	30.7 · **68.6** · 100.0
	PriorNet	99.8 · **99.8** · 99.8	99.4 · **99.8** · 99.9	98.4 · **99.7** · 99.9	49.8 · **92.7** · 99.9	31.3 · **76.6** · 99.8	30.7 · **71.8** · 100.0
	DDNet	98.5 · **98.6** · 98.7	95.0 · **97.6** · 98.9	74.4 · **91.9** · 98.2	31.4 · **52.0** · 98.5	30.7 · **51.8** · 100.0	30.7 · **40.2** · 100.0
	EvNet	85.7 · **87.5** · 89.2	69.0 · **90.4** · 97.7	42.5 · **90.2** · 99.6	30.7 · **70.1** · 100.0	30.7 · **50.0** · 100.0	30.7 · **43.9** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	84.4 · **96.3** · 99.4	50.6 · **89.3** · 99.5	30.9 · **46.3** · 99.4	30.7 · **46.3** · 100.0	30.7 · **63.3** · 100.0
	PriorNet	-	99.7 · **99.9** · 100.0	98.7 · **99.8** · 100.0	84.1 · **99.2** · 100.0	30.7 · **84.6** · 100.0	30.7 · **68.1** · 100.0
	DDNet	-	93.6 · **96.9** · 98.5	71.0 · **89.0** · 96.9	32.3 · **50.4** · 99.0	30.7 · **51.1** · 100.0	30.7 · **54.1** · 100.0
	EvNet	-	58.2 · **84.5** · 94.3	40.9 · **87.2** · 99.2	30.7 · **59.3** · 100.0	30.7 · **39.7** · 100.0	30.7 · **52.7** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	58.6 · **96.1** · 99.3	59.9 · **96.2** · 99.9	31.2 · **47.6** · 95.5	30.7 · **41.8** · 100.0	30.7 · **55.4** · 100.0
	PriorNet	-	99.9 · **100.0** · 100.0	96.6 · **99.2** · 99.9	50.3 · **97.1** · 100.0	31.7 · **89.7** · 100.0	30.7 · **81.8** · 100.0
	DDNet	-	95.0 · **97.5** · 98.8	80.5 · **94.0** · 98.6	31.7 · **55.6** · 98.6	30.7 · **52.0** · 100.0	30.7 · **49.5** · 100.0
	EvNet	-	66.5 · **91.4** · 98.1	48.5 · **84.5** · 97.6	30.8 · **49.3** · 99.9	30.7 · **37.3** · 100.0	30.8 · **62.0** · 100.0
					OOD-Attack		
Smoothed models	**PostNet**	59.2 · **91.3** · 97.7	57.9 · **97.2** · 99.6	61.4 · **93.8** · 99.6	31.5 · **59.1** · 99.5	30.7 · **52.4** · 100.0	30.7 · **53.9** · 100.0
	PriorNet	99.7 · **99.8** · 99.8	99.4 · **99.8** · 99.9	98.3 · **99.7** · 100.0	60.4 · **96.6** · 100.0	32.8 · **88.2** · 99.9	30.7 · **86.1** · 100.0
	DDNet	98.4 · **98.5** · 98.7	94.3 · **97.2** · 98.7	72.2 · **90.6** · 97.8	31.6 · **52.2** · 98.1	30.7 · **51.8** · 100.0	30.7 · **38.5** · 100.0
	EvNet	83.9 · **85.7** · 88.0	63.6 · **88.6** · 97.9	40.1 · **87.6** · 99.6	30.8 · **69.2** · 100.0	30.7 · **43.5** · 100.0	30.7 · **37.4** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	84.4 · **96.2** · 99.4	49.7 · **89.1** · 99.5	30.9 · **45.6** · 99.3	30.7 · **46.2** · 100.0	30.7 · **68.1** · 100.0
	PriorNet	-	99.7 · **99.9** · 100.0	98.7 · **99.8** · 100.0	86.3 · **99.4** · 100.0	30.9 · **91.9** · 100.0	30.7 · **77.5** · 100.0
	DDNet	-	93.9 · **97.0** · 98.6	72.1 · **89.5** · 97.0	33.0 · **52.3** · 98.8	30.7 · **51.5** · 100.0	30.7 · **60.4** · 100.0
	EvNet	-	59.4 · **85.6** · 94.6	40.7 · **86.7** · 99.2	30.7 · **57.3** · 100.0	30.7 · **39.4** · 100.0	30.7 · **49.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	55.8 · **96.1** · 99.3	58.4 · **95.7** · 99.8	31.1 · **44.6** · 93.3	30.7 · **41.4** · 100.0	30.7 · **50.1** · 100.0
	PriorNet	-	99.9 · **100.0** · 100.0	96.9 · **99.3** · 99.9	60.3 · **98.2** · 100.0	33.0 · **93.5** · 100.0	30.7 · **87.8** · 100.0
	DDNet	-	95.3 · **97.6** · 98.9	82.3 · **94.5** · 98.7	32.1 · **56.3** · 98.5	30.7 · **48.9** · 100.0	30.7 · **43.4** · 100.0
	EvNet	-	65.3 · **90.3** · 97.9	46.9 · **83.1** · 97.3	30.8 · **48.8** · 99.9	30.7 · **36.6** · 100.0	30.8 · **60.7** · 100.0

Table 51. OOD detection based on differential entropy under FGSM uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on Sensorless. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model)..

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
					ID-Attack		
Smoothed models	**PostNet**	49.3 · **90.4** · 99.8	30.7 · **50.3** · 100.0	30.7 · **36.6** · 100.0	49.1 · **50.0** · 74.9	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	31.2 · **39.0** · 66.9	30.7 · **40.1** · 100.0	30.7 · **48.2** · 100.0	30.7 · **54.2** · 100.0	30.7 · **46.3** · 100.0	30.7 · **47.6** · 100.0
	DDNet	31.0 · **31.5** · 32.7	30.7 · **31.2** · 100.0	30.7 · **35.3** · 100.0	30.7 · **55.7** · 100.0	30.7 · **42.4** · 100.0	30.7 · **40.4** · 100.0
	EvNet	33.6 · **55.1** · 91.3	30.7 · **39.1** · 100.0	30.7 · **37.1** · 100.0	30.7 · **35.4** · 100.0	30.8 · **52.1** · 100.0	32.5 · **50.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	30.7 · **60.8** · 100.0	30.7 · **40.7** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **31.3** · 100.0	30.7 · **32.9** · 100.0	30.7 · **40.1** · 100.0	30.8 · **31.1** · 100.0	38.1 · **91.0** · 100.0
	DDNet	-	30.7 · **34.3** · 100.0	30.7 · **33.9** · 100.0	30.7 · **38.2** · 100.0	30.7 · **63.6** · 100.0	30.7 · **41.8** · 100.0
	EvNet	-	30.8 · **41.0** · 100.0	30.7 · **34.2** · 100.0	30.7 · **38.0** · 100.0	30.7 · **39.0** · 100.0	30.7 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.7 · **46.1** · 100.0	30.7 · **46.8** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **36.5** · 100.0	30.7 · **34.4** · 100.0	30.7 · **77.8** · 100.0	30.8 · **53.0** · 100.0	30.7 · **39.2** · 100.0
	DDNet	-	30.7 · **36.0** · 100.0	30.7 · **37.7** · 100.0	30.7 · **41.0** · 100.0	30.7 · **42.3** · 100.0	30.7 · **39.0** · 100.0
	EvNet	-	30.7 · **31.3** · 100.0	30.7 · **43.3** · 100.0	30.7 · **36.3** · 100.0	30.7 · **43.2** · 100.0	30.7 · **50.0** · 100.0
					OOD-Attack		
Smoothed models	**PostNet**	49.3 · **90.4** · 99.8	30.8 · **75.3** · 100.0	30.7 · **68.5** · 100.0	46.1 · **50.0** · 74.8	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	31.2 · **38.9** · 67.0	30.7 · **34.1** · 100.0	30.7 · **35.7** · 100.0	30.7 · **35.0** · 100.0	30.7 · **77.6** · 100.0	30.8 · **95.3** · 100.0
	DDNet	31.0 · **31.5** · 32.7	30.7 · **30.8** · 100.0	30.7 · **33.1** · 100.0	30.7 · **65.7** · 100.0	30.7 · **71.8** · 100.0	30.7 · **71.5** · 100.0
	EvNet	33.6 · **55.2** · 91.4	30.7 · **64.7** · 100.0	30.7 · **69.6** · 100.0	30.7 · **78.9** · 100.0	30.7 · **67.2** · 100.0	32.9 · **50.0** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	30.7 · **86.0** · 100.0	30.7 · **86.6** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **31.0** · 99.9	30.7 · **31.2** · 98.9	30.7 · **30.7** · 100.0	30.8 · **30.7** · 100.0	36.1 · **35.3** · 100.0
	DDNet	-	30.7 · **37.2** · 100.0	30.7 · **31.1** · 100.0	30.7 · **37.1** · 100.0	30.7 · **50.5** · 100.0	30.7 · **84.6** · 100.0
	EvNet	-	30.8 · **82.5** · 100.0	30.7 · **51.7** · 100.0	30.7 · **91.5** · 100.0	30.7 · **70.0** · 100.0	30.9 · **50.0** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.7 · **78.5** · 100.0	30.7 · **67.1** · 100.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.7 · **35.8** · 100.0	30.7 · **30.7** · 100.0	30.7 · **39.0** · 100.0	30.7 · **58.5** · 100.0	30.7 · **100.0** · 100.0
	DDNet	-	30.7 · **40.8** · 100.0	30.7 · **33.1** · 100.0	30.7 · **30.8** · 100.0	30.7 · **34.3** · 100.0	30.7 · **35.2** · 100.0
	EvNet	-	30.7 · **32.7** · 100.0	30.8 · **50.2** · 100.0	30.7 · **99.6** · 100.0	30.7 · **58.7** · 100.0	30.7 · **50.0** · 100.0

Table 52. OOD detection based on differential entropy under FGSM uncertainty attacks against differential entropy on ID data and OOD data. Smoothed DBU models on Segment. Column format: guaranteed lowest performance · empirical performance · guaranteed highest performance (blue: normally/adversarially trained smooth classifier is more robust than the base model).

	Att. Rad.	0.0	0.1	0.2	0.5	1.0	2.0
				ID-Attack			
Smoothed models	**PostNet**	99.6 · **99.9** · 99.9	33.1 · **78.8** · 100.0	30.8 · **46.2** · 100.0	30.8 · **34.2** · 100.0	30.8 · **41.4** · 100.0	41.5 · **50.0** · 50.2
	PriorNet	30.9 · **31.0** · 31.4	30.8 · **30.8** · 39.3	30.8 · **30.8** · 94.7	30.8 · **41.2** · 100.0	30.8 · **92.7** · 100.0	30.8 · **79.9** · 100.0
	DDNet	30.8 · **30.8** · 30.8	30.8 · **30.8** · 31.8	30.8 · **30.8** · 66.8	30.8 · **30.8** · 100.0	30.8 · **32.6** · 100.0	30.8 · **38.2** · 100.0
	EvNet	94.9 · **97.2** · 98.2	31.0 · **73.1** · 100.0	30.8 · **72.3** · 100.0	30.8 · **57.1** · 100.0	30.8 · **63.3** · 100.0	30.8 · **49.6** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	31.0 · **62.9** · 100.0	30.8 · **47.1** · 100.0	30.8 · **90.0** · 100.0	30.8 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **30.8** · 43.5	30.8 · **30.8** · 32.5	30.8 · **30.9** · 100.0	30.8 · **30.9** · 100.0	30.8 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 30.8	30.8 · **30.8** · 76.1	30.8 · **30.9** · 100.0	30.8 · **34.8** · 100.0	30.8 · **53.0** · 100.0
	EvNet	-	35.5 · **93.5** · 100.0	30.8 · **31.8** · 100.0	30.8 · **48.7** · 100.0	30.8 · **93.8** · 100.0	30.8 · **63.7** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.8 · **47.5** · 100.0	30.8 · **37.5** · 100.0	30.8 · **92.9** · 100.0	41.1 · **50.0** · 97.3	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **31.1** · 60.8	30.8 · **30.8** · 32.3	30.8 · **30.8** · 90.3	30.8 · **30.8** · 100.0	30.8 · **36.3** · 100.0
	DDNet	-	30.8 · **30.8** · 31.0	30.8 · **30.8** · 66.8	30.8 · **30.8** · 100.0	30.8 · **31.2** · 100.0	30.8 · **57.2** · 100.0
	EvNet	-	30.9 · **80.3** · 100.0	30.8 · **78.1** · 100.0	30.8 · **99.4** · 100.0	30.8 · **97.7** · 100.0	30.8 · **41.5** · 100.0
				OOD-Attack			
Smoothed models	**PostNet**	99.6 · **99.9** · 99.9	31.2 · **94.3** · 100.0	30.8 · **44.8** · 100.0	30.8 · **36.8** · 100.0	30.8 · **39.9** · 100.0	44.3 · **50.0** · 50.0
	PriorNet	30.9 · **31.0** · 31.4	30.8 · **30.8** · 42.0	30.8 · **30.8** · 80.4	30.8 · **30.8** · 100.0	30.8 · **37.5** · 100.0	30.8 · **94.9** · 100.0
	DDNet	30.8 · **30.8** · 30.8	30.8 · **30.8** · 31.5	30.8 · **30.8** · 48.0	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	EvNet	94.9 · **97.2** · 98.3	31.3 · **92.1** · 100.0	30.8 · **90.8** · 100.0	30.8 · **89.6** · 100.0	30.8 · **89.8** · 100.0	30.8 · **87.3** · 100.0
Smoothed + adv. w. label attacks	**PostNet**	-	30.8 · **85.3** · 100.0	30.8 · **85.9** · 100.0	30.8 · **78.8** · 100.0	30.9 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **30.8** · 45.0	30.8 · **30.8** · 32.1	30.8 · **30.8** · 90.3	30.8 · **30.8** · 100.0	31.0 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 30.8	30.8 · **30.8** · 64.9	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **79.4** · 100.0
	EvNet	-	35.4 · **95.0** · 100.0	30.8 · **35.2** · 100.0	30.8 · **51.9** · 100.0	30.8 · **80.0** · 100.0	30.8 · **99.9** · 100.0
Smoothed + adv. w. uncert. attacks	**PostNet**	-	30.8 · **63.4** · 100.0	30.8 · **31.7** · 100.0	30.8 · **98.4** · 100.0	33.2 · **50.0** · 100.0	50.0 · **50.0** · 50.0
	PriorNet	-	30.8 · **31.1** · 58.0	30.8 · **30.8** · 34.1	30.8 · **30.8** · 66.8	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	DDNet	-	30.8 · **30.8** · 31.2	30.8 · **30.8** · 61.5	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0	30.8 · **30.8** · 100.0
	EvNet	-	31.0 · **89.0** · 100.0	30.8 · **96.2** · 100.0	30.8 · **99.6** · 100.0	30.8 · **99.6** · 100.0	30.8 · **69.7** · 100.0

6.5. Visualization of differential entropy distributions on ID data and OOD data

The following Figures visualize the differential entropy distribution for ID data and OOD data for all models with standard training. We used label attacks and uncertainty attacks for CIFAR10 and MNIST. Thus, they show how well the DBU models separate on clean and perturbed ID data and OOD data.

Figures 4 and 5 visualizes the differential entropy distribution of ID data and OOD data under label attacks. On CIFAR10, PriorNet and DDNet can barely distinguish between clean ID and OOD data. We observe a better ID/OOD distinction for PostNet and EvNet for clean data. However, we do not observe for any model an increase of the uncertainty estimates on label attacked data. Even worse, PostNet, PriorNet and DDNet seem to assign higher confidence on class label attacks. On MNIST, models show a slightly better behavior. They are capable to assign a higher uncertainty to label attacks up to some attack radius.

Figures 6, 7, 8 and 9 visualizes the differential entropy distribution of ID data and OOD data under uncertainty attacks. For both CIFAR10 and MNIST data sets, we observed that uncertainty estimations of all models can be manipulated. That is, OOD uncertainty attacks can shift the OOD uncertainty distribution to more certain predictions, and ID uncertainty attacks can shift the ID uncertainty distribution to less certain predictions.

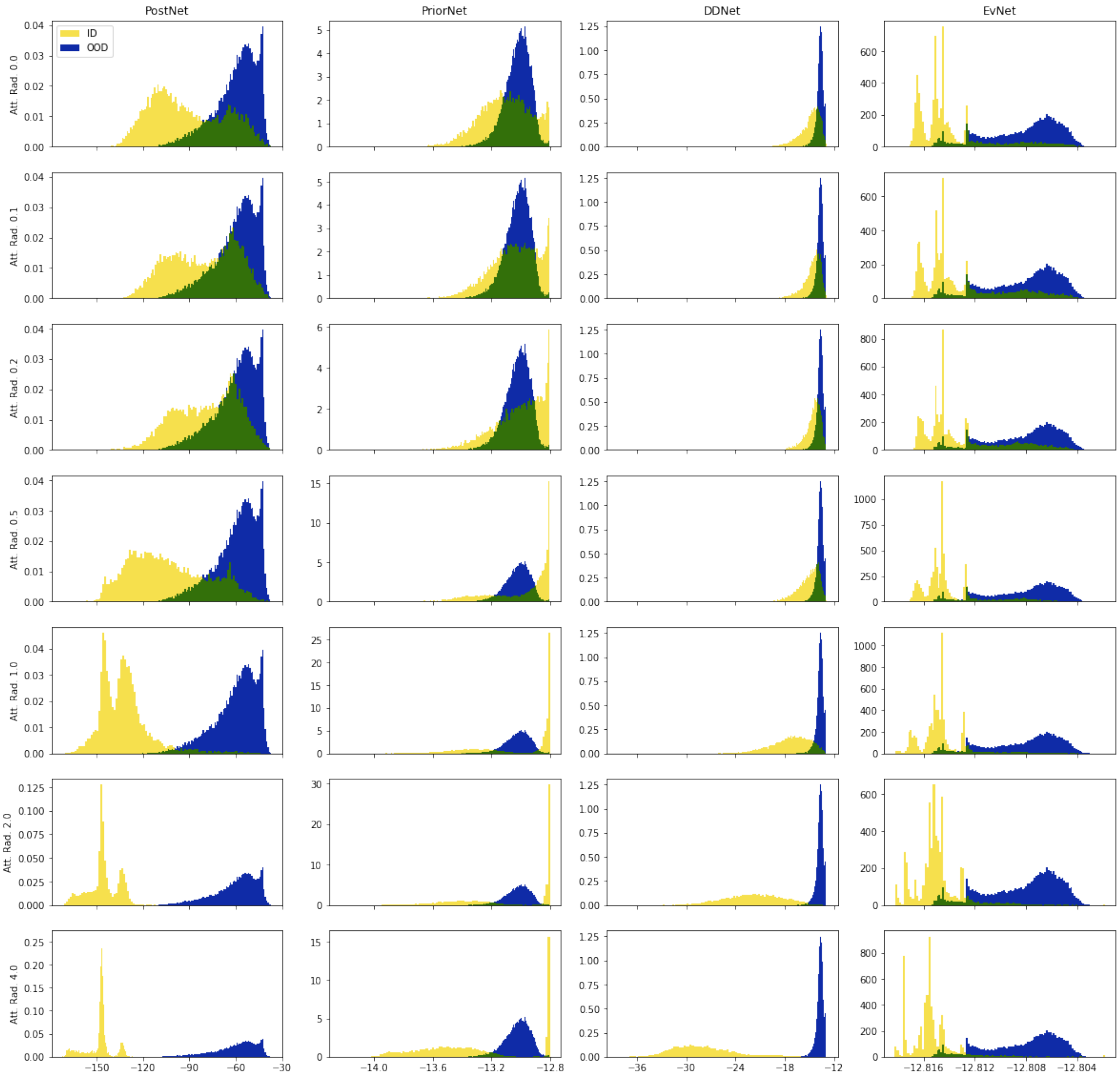

Figure 4. Visualization of the differential entropy distribution of ID data (CIFAR10) and OOD data (SVHN) under label attack. The first row corresponds to no attack. The other rows correspond do increasingly stronger attack strength.

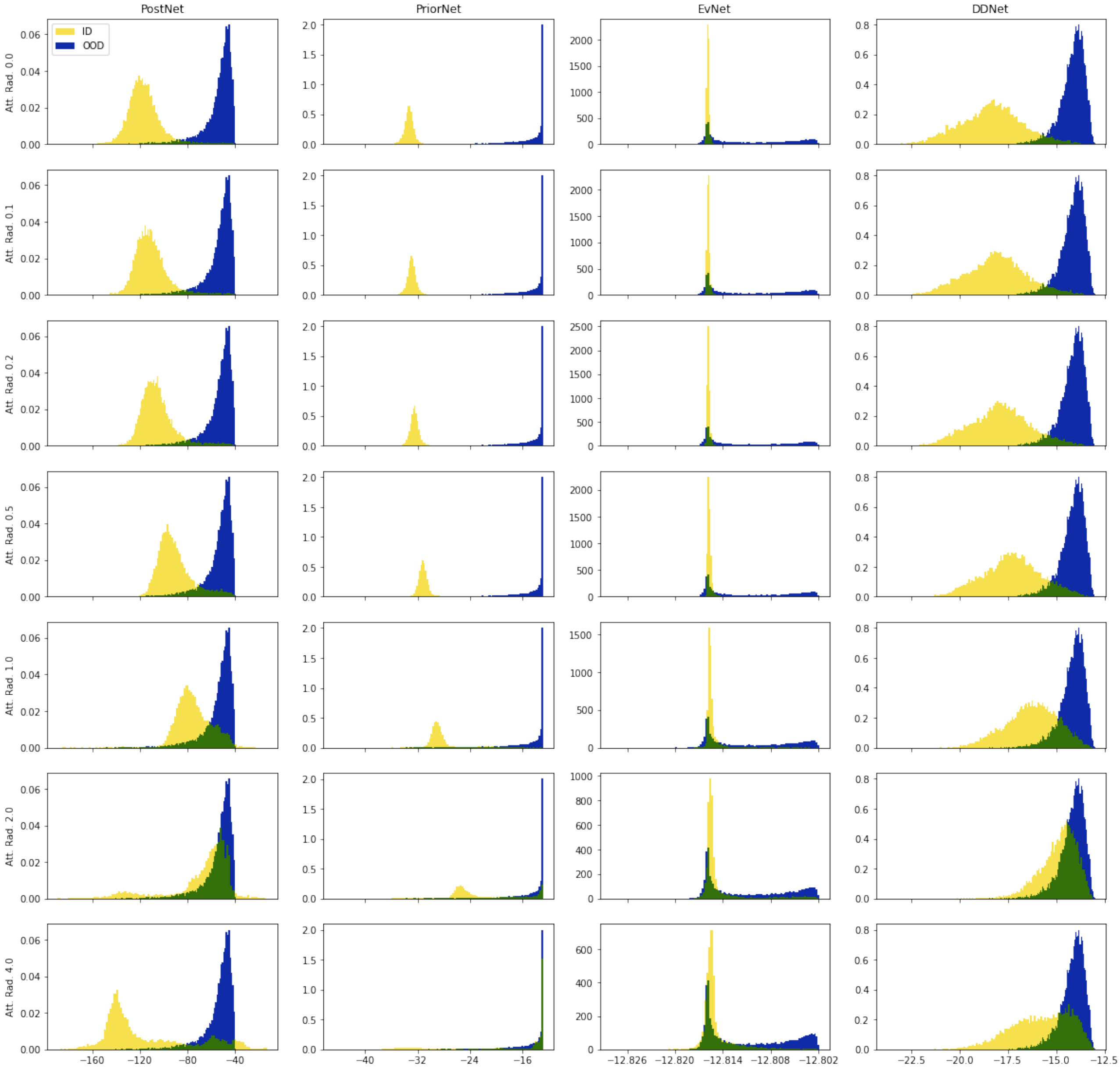

Figure 5. Visualization of the differential entropy distribution of ID data (MNIST) and OOD data (KMNIST) under label attack. The first row corresponds to no attack. The other rows correspond do increasingly stronger attack strength.

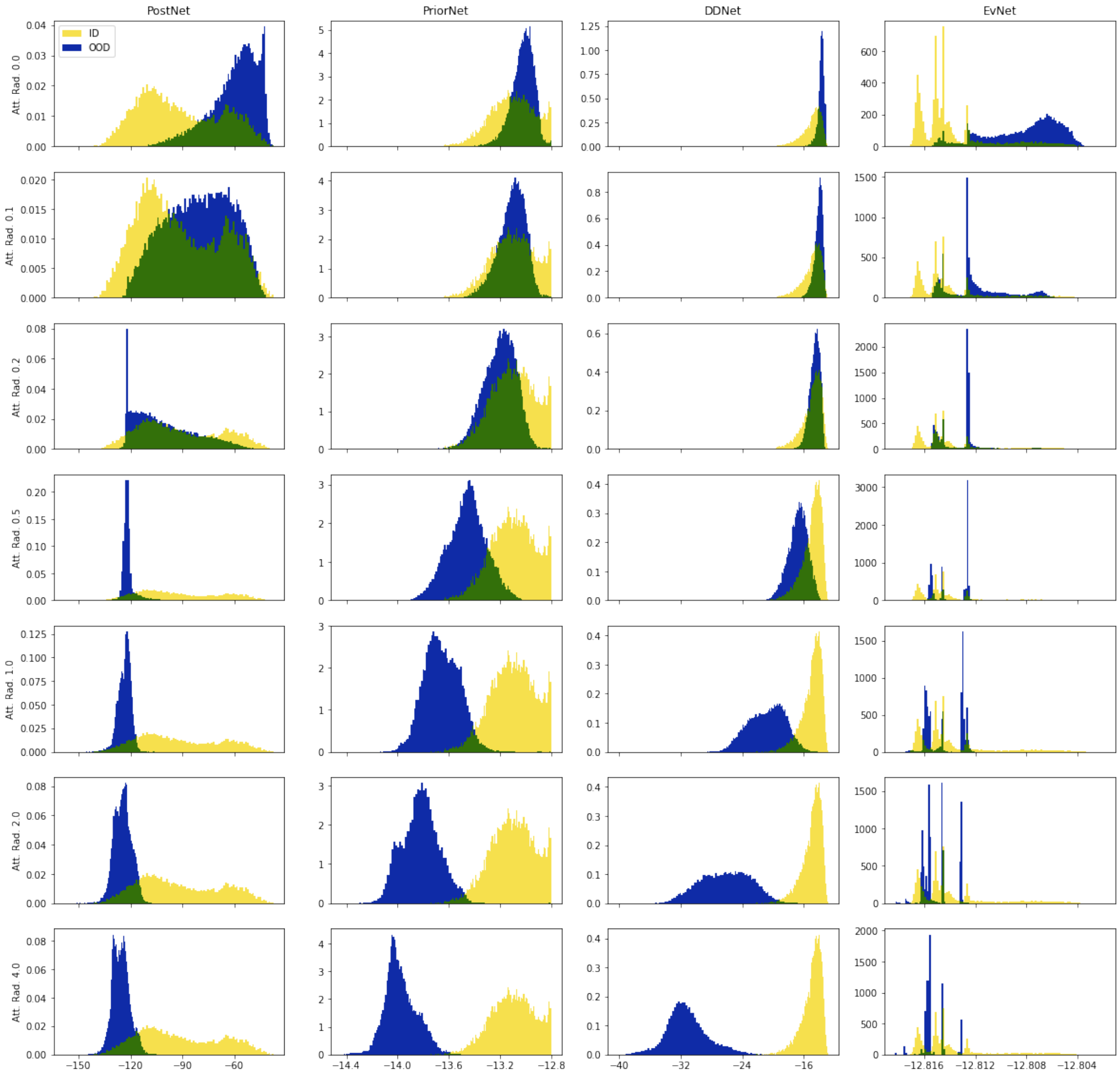

Figure 6. Visualization of the differential entropy distribution of ID data (CIFAR10) and OOD data (SVHN) under OOD uncertainty attack. The first row corresponds to no attack. The other rows correspond do increasingly stronger attack strength.

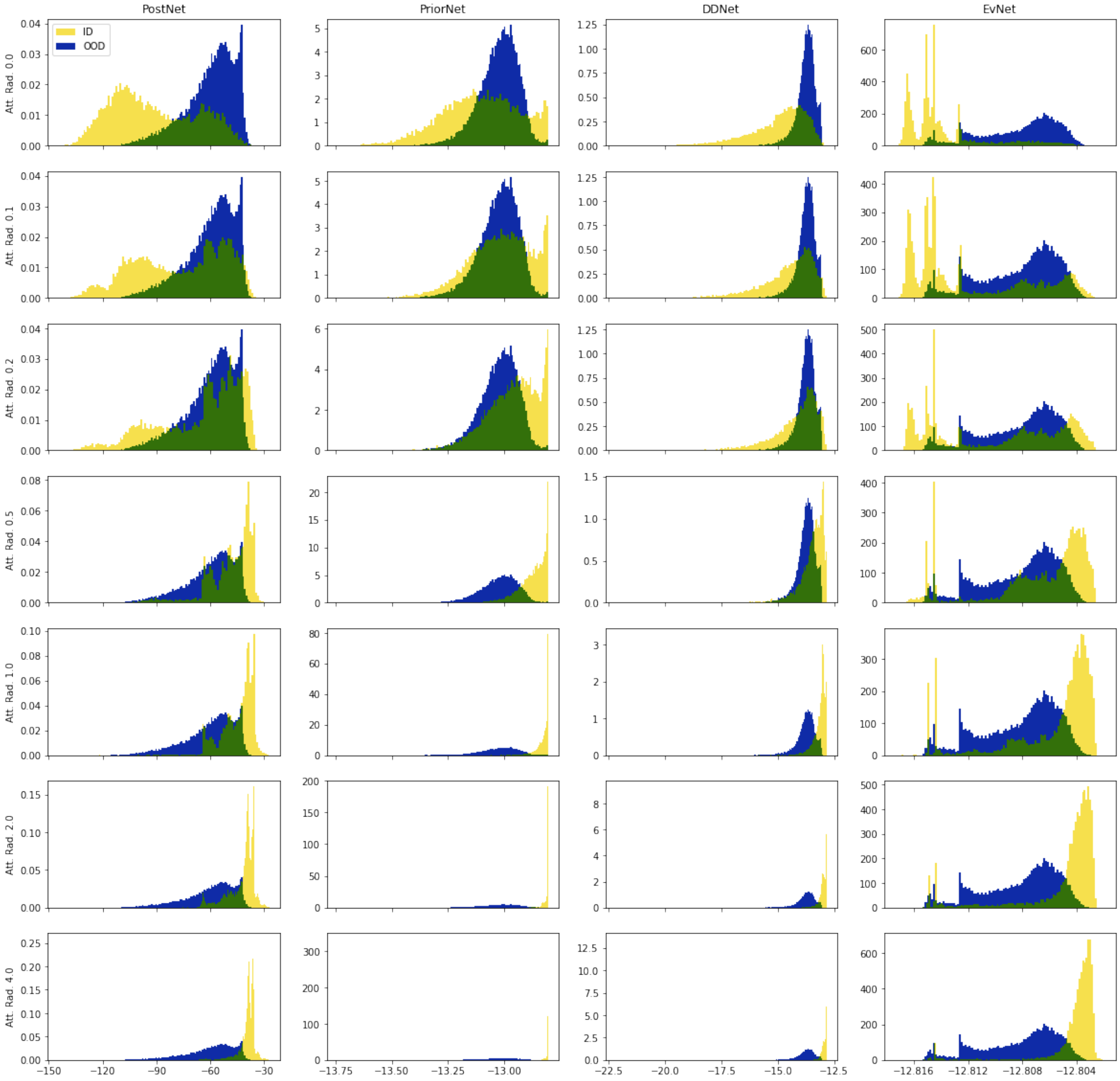

Figure 7. Visualization of the differential entropy distribution of ID data (CIFAR10) and OOD data (SVHN) under ID uncertainty attack. The first row corresponds to no attack. The other rows correspond do increasingly stronger attack strength.

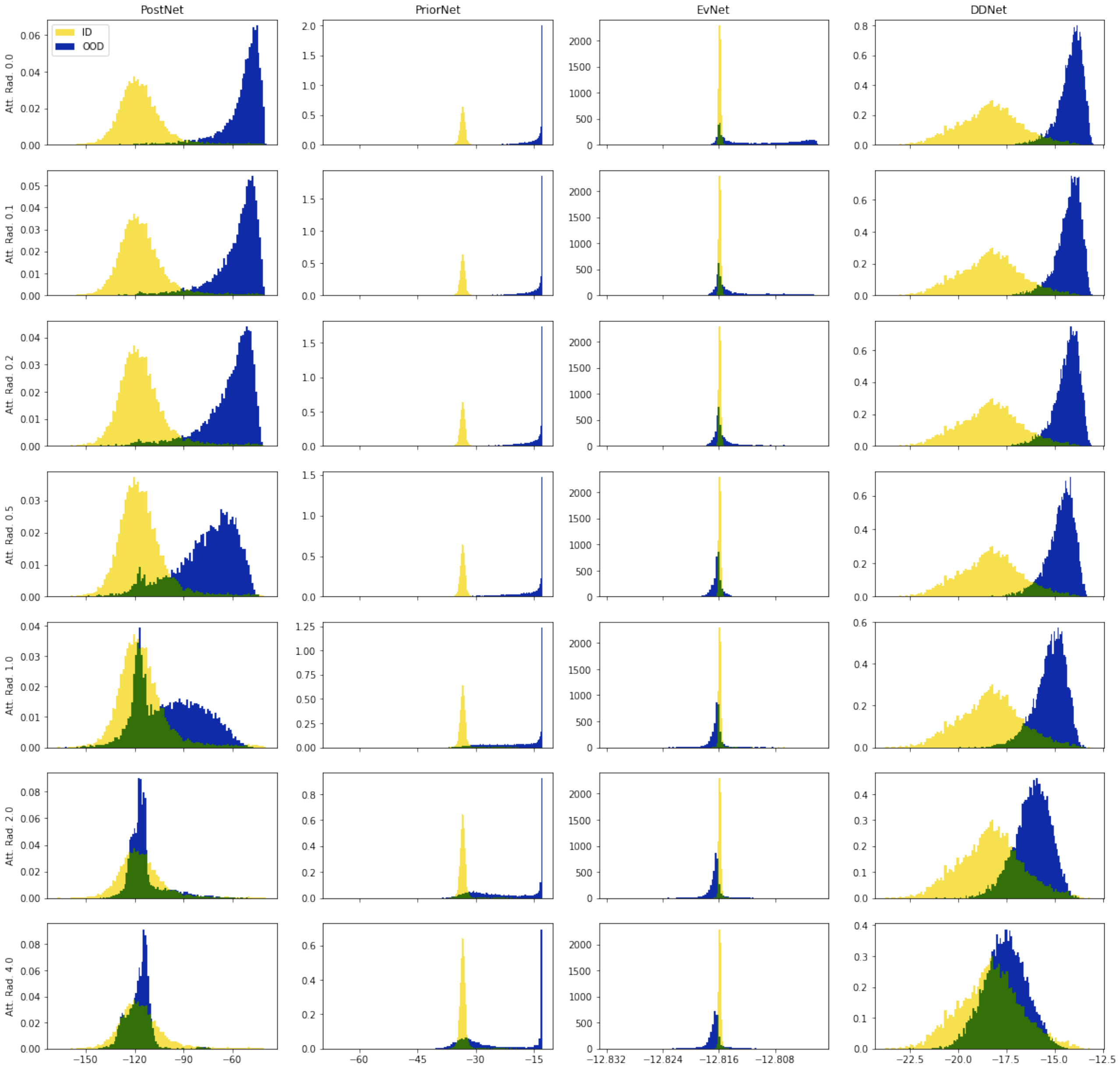

Figure 8. Visualization of the differential entropy distribution of ID data (MNIST) and OOD data (KMNIST) under OOD uncertainty attack. The first row corresponds to no attack. The other rows correspond do increasingly stronger attack strength.

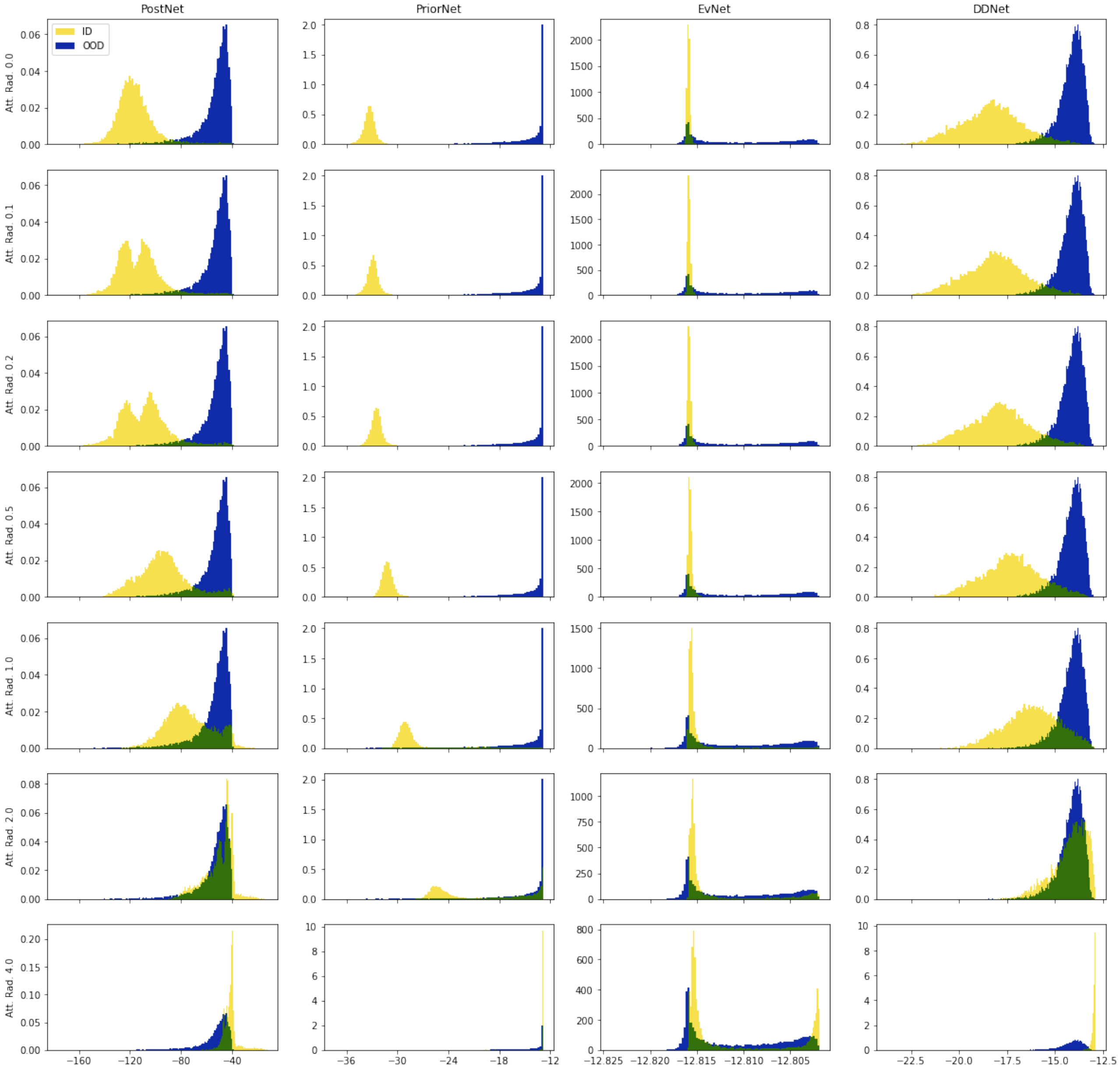

Figure 9. Visualization of the differential entropy distribution of ID data (MNIST) and OOD data (KMNIST) under ID uncertainty attack. The first row corresponds to no attack. The other rows correspond do increasingly stronger attack strength.